2002

An investigation of time series and case-crossover analyses of air pollution and asthma hospital admission data for children in Toronto (Ontario).

Abby Leigh Livingston
University of Windsor

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

AN INVESTIGATION OF TIME SERIES AND CASE-CROSSOVER ANALYSES OF AIR POLLUTION AND ASTHMA HOSPITAL ADMISSION DATA FOR CHILDREN IN TORONTO

BY

ABBY LIVINGSTON

A Thesis

Submitted to the Faculty of Graduate Studies and Research

Through the Department of Mathematics and Statistics

in Partial Fulfillment of the Requirements for

the Degree of Master of Science at the

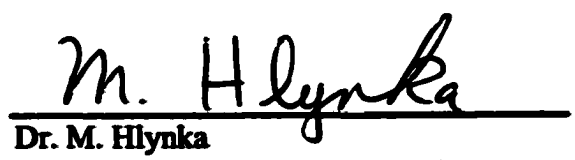
University of Windsor

Windsor, Ontario, Canada
2002

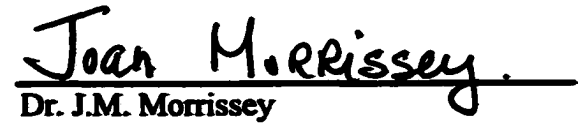
973133

Approved by:

Dr. K. Fung, Supervisor
Department of Mathematics and Statistics



Dr. M. Hlynka
Department of Mathematics and Statistics



Dr. J.M. Morrissey
Department of Computer Science

Abstract

Air pollution has been a popular topic of study over the years. It causes great harm to our environment (global warming) not to mention our health (cancer, heart disease, respiratory disease, *etc.*). Many people have investigated the damaging relationship between air pollution and mortality and morbidity, using different methods along the way.

The different methods yield results that are not directly comparable with one another because the methods use different strategies. Air pollution data and hospital admissions data for asthma patients aged six to twelve in the Toronto area from January 1, 1981 to December 31, 1993 were gathered and analyzed under a variety of time series and case-crossover designs. The lack of consistency in the results among the techniques led us to perform a simulation in order to choose the most accurate method to analyze this Toronto data.

While the time series approach produced fairly accurate results, the bidirectional case-crossover using the exact method of approximation was the overall best technique of analysis.

Acknowledgements

I wish to express my sincere thanks to my supervisor Dr. K. Fung for her guidance in this statistics field. More importantly, I wish to thank her for showing me the amazing field of biostatistics. I would also like to thank Drs. M. Hlynka and J. Morrissey for serving on my advisory committee.

I am also grateful to the Department of Mathematics and Statistics for providing financial support in terms of teaching assistantships, graduate assistantships and scholarships throughout my years of study. The research assistantship from Dr. K. Fung is also appreciated.

I would also like to thank all of my great friends, new and old. Without them my time as a graduate student would not have been as enjoyable.

Finally, I express a very special gratitude for my entire family, especially my parents, for their love, support and encouragement during my studies.

Contents

Chapter 1

Introduction

Air pollution has been a major concern in North America for many years. With the increase in the use of automobiles and rapid factory expansion, air pollution has become an even bigger problem. Not only is it damaging to our ozone layer and natural resources, such as forests and the Great Lakes, it also endangers our health as Schwartz *et al.* (1996) and Koeing (1999) show us.

Air pollution can be broken down into various gases and particulates. In turn, each of these can be studied independently or jointly with one another. The main gaseous constituents are carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), and ozone (O_3). These gases are most harmful to young children, the elderly and those suffering from respiratory or cardiac diseases as shown in Schwartz and Morris (1995) as well as in Burnett *et al.* (1999), but healthy people are also susceptible to their effects.

Particulate matter is the airborne solid or liquid particles of smoke, ash and dust. It comes mainly from vehicle emissions and can cause extensive lung damage, as the particles are able to penetrate lung tissue. Particulate matter is classified by the size of the particles. In this thesis, PM_{10} refers to particles measuring less than 10 microns in diameter and $PM_{2.5}$ refers to particles measuring less than 2.5 microns in diameter.

Burnett *et al.* (1995) studied the effects of particulate matter in great detail and found it to have a significant association with cardiorespiratory admissions.

Carbon monoxide, both odorless and colourless, enters our air via automobile exhaust. Once in our lungs, it enters the bloodstream, reducing the oxygen delivered to bodily tissues. The lack of oxygen weakens our heart, thus reducing the amount of blood being pumped through the body.

Nitrogen dioxide is the brown gas that can be seen emanating from industrial smoke stacks. By impairing the body's respiratory defenses, NO_2 increases one's vulnerability to various infections. The other major difficulty with NO_2 is that it helps to create ozone.

The byproduct ozone found in air pollution is not to be confused with the naturally occurring ozone found in our atmosphere. While natural ozone is beneficial to us, as it filters ultraviolet radiation, the ozone resulting from a reaction of NO_2 with hydrocarbons (found in vehicular exhaust) in sunlight is extremely harmful to us. Approximately 90% of the ozone inhaled by humans is never exhaled, and remains in the lungs.

Sulfur dioxide, colourless and odorless at low levels, yet pungent at high, is produced through industrial processes. It impairs respiratory function for everyone and is especially aggravating to those suffering from asthma. Because SO_2 decreases the lungs' ability to remove foreign particles, it is extremely harmful in connection with high levels of particulate matter.

We know that air pollution affects our health, but is it a significant factor or is it just another of many contributors such as lack of exercise or poor diet. There have been a

tremendous number of studies performed trying to link air pollution to various health problems such as Burnett *et al.* (1994), Lee and Schwartz (1999), and Neas *et al.* (1999). Many are convinced that respiratory and cardiac illnesses are significantly worsened and may even be triggered by the air we breathe.

London, England in December of 1952 is a prime example of this theory. At this time there were increased particle and sulfur dioxide concentrations in the air, as well as increased counts for daily deaths and hospital admissions. Both respiratory and cardiac causes of death and admission were substantially higher than usual. Studies, in particular the aforementioned done by Schwartz and Morris (1995), have shown a positive association between heightened air pollution and the escalation in deaths in London during this time. Similarly, Burnett and Krewski (1994) studied how pollutants affect hospital admissions and found a definite correlation with data from 164 Ontario hospitals during the 1980's. Although respiratory complications have been positively correlated with air pollution, Schwartz and Morris (1995) studied hospital admissions in Detroit, Michigan and found air pollution to be associated with cardiovascular disease as well.

In Ontario during the mid nineties, (1995/96), the government made severe cuts to the health care system. This created numerous problems such as fewer hospital beds and longer lines in emergency rooms. It would be a great help to the system if there were a way to lower the number of respiratory admissions to the hospital. By positively identifying the link between pollution and hospital admissions we can focus on cleaning the air, consequently lowering hospital admissions.

Chapter 2 of this thesis investigates the associations between asthma hospitalizations of children aged six to twelve years in the Toronto area and gaseous

pollutants or particulate matter. These results are inconclusive since we cannot tell which method of analysis is the most accurate. Using the $PM_{2.5}$ Toronto data, a simulation was performed in Chapter 3 to compare each method of analysis and decide upon the most precise one. Once the decision has been made, we revisit the results of Chapter 2 to make a final conclusion. This conclusion, along with limitations to this investigation are presented in Chapter 4. The main statistical topics used in this thesis such as Cox proportional hazards model, the conditional likelihood, logistic regression, autocorrelations and the generalized additive model are elaborated upon in Chapter 5.

Chapter 2

Analysis of Asthma Hospitalizations and Air Pollution in Toronto, Ontario

2.1 Source of Data and Description of Data

The Ontario Ministry of Health maintains data on hospital admissions across the province. The daily admission counts for asthma conditions in children aged six to twelve in the Toronto area were obtained and used in this analysis. Weather conditions for this time period (humidity, maximum temperature and minimum temperature) were acquired from the Pearson International Airport in Toronto. The Ontario Ministry of Environment and Energy (OMEE) provided the daily concentrations of CO, NO_2, SO_2, O_3, $PM_{2.5}$, and PM_{10} as were recorded from four monitoring stations throughout Toronto. All of the data that were gathered span from January 1, 1981 through December 31, 1993.

Summaries of this Toronto data set are provided in Table 1. The mean number of daily asthma hospitalizations for girls, boys and the two combined (ages 6 through 12) are 0.57, 0.97, and 1.54 respectively. Female admissions reached a daily maximum of 5, whereas the counts for males and males/females combined peaked at 11 each.

Of the entire period, the highest daily maximum temperature was 37.6 °C, and the lowest maximum temperature was recorded as -19.8 °C. Similarly, for the minimum daily temperature, 24.3 °C and -31.3 °C were the highest and lowest values respectively.

The mean daily maximum and minimum temperatures were 12.68 °C and 2.63 °C respectively. The relative humidity has a mean of 73.51%, where 99% was the highest daily value recorded and 35% was the lowest.

Canada and the United States of America have National Ambient Air Quality Standards (NAAQS) as set by the Environmental Protection Agency (EPA). By monitoring air pollution, these standards were set to protect our health, paying close attention to children, the elderly and those with respiratory conditions. Carbon monoxide, nitrogen dioxide, ozone, lead, particulate matter and sulfur dioxide are the six principal pollutants for which standards have been set. Parts per million/billion by volume (ppm or ppb) and micrograms per cubic meter of air ($\mu g / m^3$) are the units of measure for these standards.

In this period, 1981 to 1993, each of the pollutants' daily means were well under the NAAQS, but there were a few days when $PM_{2.5}$ (particulate matter measuring 2.5 micrometers or less in diameter) was actually above the standard level of 65 $\mu g / m^3$ over a 24hr period. The range of values over the period for each pollutant was quite large. Nitrogen dioxide, for example, had a low of 3 ppb and a high of 82 ppb.

To study the relationship between personal exposure to air pollutants and asthma hospitalizations, we assume that one's personal exposure is equal to the ambient air pollution level from the centrally sited outdoor monitors. Because of this, measurement error may exist in the data.

2.2 Methods

The means and inter-quartile ranges for each of the gases (NO_2, SO_2, O_3, and CO) and particulate matters (PM_{10}, and $PM_{2.5}$) were calculated for the Toronto data. Each

pollutant was studied independently under a variety of time series and case-crossover designs.

2.2.1 Time Series Analysis

The Toronto data set was analyzed using both co-adjusted and pre-adjusted time series analysis techniques. The co-adjusted method examines temporal trends and air pollution predictors together in a generalized additive model (GAM). The pre-adjustment approach removes temporal trends from the health and air pollution time series prior to linking them together.

Temporal trends were removed by selecting a nonparametric smoothed function of day of study (LOESS) with the span chosen such that autocorrelation in the residuals is minimal. LOESS is a generalization of a weighted moving average. The smoother is characterized by defining a window of observations with fixed span about a specific date. Greater weights are put on days closer to the center of the window and decrease towards zero at the boundaries within the smoothing function.

The best span for a pollutant is the one that yields the smallest positive residuals, and the fewest negative residuals. The longer the lags become, the smaller (closer to 0) the coefficients become. Appropriate spans for each pollutant were determined by analyzing the auto-correlation functions (ACF) of the data. Correlograms, or ACF plots, estimate the correlation between observations separated by various time units. Although plotting the data against time may show distinct trends, the ACF plots are easily plotted using S-Plus and less obvious serial correlations can be seen from them. Using SO_2 as an example, the spans of one year, half a year, three months, and two months were considered. After comparing the graphs in Figure 1, it is clear that half a year has the

best fit and has the smallest and fewest residuals. Figure 2 shows ACF plots for each of our six pollutants using the spanning width which best fits that particular pollutant.

The daily admission counts for each day of the week is another trend that requires some investigation. Not every day of the week has the same probability of admitting someone to be treated for asthma. Because of this, a day of the week indicator variable, D_t, was also considered as a covariate. D_t is a time series consisting of a repetition of seven values representing the ratio of the average number of admissions on each of the seven days of the week to the average daily admission rate. Starting with Monday these seven values are: 1.16, 1.07, 1.00, 1.01, 1.01, 0.84, and 0.89. (Note that D_t reflects the fact that there are more admissions on Mondays and less on the weekends.)

A generalized additive model (GAM) was fit to the Toronto admissions data with each air pollutant, adjusting for day of the week and a LOESS smoothing function (with the pollutant's appropriate span) as covariates. This model fits nonparametric functions to estimate the relationship between air pollution levels and asthma hospitalization. Quasi-likelihood estimation was used to take into consideration over- or under-dispersion of daily hospital admission counts.

The following model was fit to the Toronto data.

$$\ln E(y_t) = \beta_0 + \beta_1 x_t + lo(time) + D_t \qquad (2.1)$$

$E(y_t)$ is the mean number of admissions to the hospital due to asthma. β_0 and β_1 are the parameters for the intercept and the pollutant being analyzed, x. Temporal trends are taken care of by *lo(time)*, which is a locally weighted regression model (LOESS) used to smooth time. The covariate, D_t, was used to account for the fact that different days of the

week have different probabilities of having admissions. All analyses performed on the data were done using S-Plus.

Many people have discovered that temporal factors are important covariates, in particular temperature and humidity (Castellagué *et al.*, 1995; Linn *et al.*, 1985; Sunyer *et al.*, 1996 and 1997). These studies showed a change in the degree of pollutant and asthma association from a straightforward analysis, to one which accounted for weather conditions.

Since temporal trends were of primary concern, the data set was reanalyzed. The second analysis was essentially the same except it used maximum temperature (*MAT*), minimum temperature (*MIT*) and relative humidity (*RH*) as additional covariates.

$$\ln E(y_t) = \beta_0 + \beta_1 x_t + lo(time) + D_t + MIT + MAT + RH \qquad (2.2)$$

2.2.2 Case-Crossover Analysis

The case-crossover model compares an individual's pollutant exposure during a case period with their own exposure during a non-case or control period. For these data, the cases are defined as hospital admission dates, and controls are the non-admission days. Maclure (1991) proposed the case-crossover method and justified reasons for and against case-crossover. Navidi (1998) modified Maclure's case-crossover technique to accommodate a bidirectional design. In recent applications, Neas *et al.* (1999) used case-crossover to analyze the effects of air pollution on mortality in Philadelphia, Pennsylvania, and Lee and Schwartz (1999) analyzed these same effects in Seoul, Korea.

Case-crossover is similar to case-control, although case-crossover is the preferred design for these data as it eliminates some of the problems that arise with the case-control design. With case-control, controls must be gathered that are characteristic of the

population from which the cases were generated. Once this has been completed, each case must be matched to a control. This often causes a bias, as it is difficult to link two individuals (one being the case, and one as the control), having the same exposure and susceptibility factors. Factors such as a case individual's age and race may be different from the control individual's, which could contribute to the bias in the results. Case-crossover uses the same individual for both the case and control so the issue of bias in the case-control design is not relevant here.

The hospital admissions data were analyzed using case-crossover in essentially two ways, uni- and bidirectional. Here, unidirectional refers to a case being compared with one control either previous to or after the date of admission. Bidirectional compares the case to two controls, one prior to and one subsequent to the case date. Conditional logistic regression models were fit to the unidirectional designs and bidirectional designs. The estimates within the bidirectional designs were obtained using three estimation methods: Breslow, Efron, and exact.

Odds ratios (OR) were calculated for each pollutant in relation to asthma hospitalization, both with and without adjustment for temperature and humidity. In this thesis odds ratios were calculated based on an increment in exposure corresponding to the inter-quartile range of each pollutant.

Since each day of the week has a different likelihood of the number of admissions it will receive, control dates were taken in weekly time periods from the cases. In this analysis, a period of two weeks between the case and control was chosen. If the time interval between the case and the control is short, the problem of possible weather or seasonal differences over the interval is taken care of. On the other hand, if the interval is

too short, there may be some autocorrelation between the exposures. A picture illustrating the two week time period is included as Figure 3.

2.3 Results

Daily hospital admissions for asthma are plotted in terms of time in Figures 4 and 5. These plots have been smoothed from their original state using the LOESS smoothing function. The original plots (Figure 6) are very hard to analyze since it is hard to tell if any trends exist due to the 'white noise' in the data. Using the LOESS smoothing function to eliminate the white noise, we produced plots that were much better for visualizing trends and cycles. Figure 4 shows the admission counts for girls, boys and the two combined, plotted against time. The peaks and valleys on all 3 plots are evenly spaced out and seem to follow a pattern. The peaks are occurring during the summer or early fall, whereas the valleys are occurring during mid winter. Thus, asthma hospitalizations are more prevalent in the warmer months than in the cooler ones. Towards the end of the study period, admission counts showed a slight decline for both males and females.

Plots pertaining to the average day of the week counts as well as the average monthly counts are illustrated in Figure 5. Mondays have the highest number of asthma admissions (mean of 1.93) whereas the rest of the week shows a steady decline with Saturdays averaging only 1.31 admissions per week. When classified by month, we see that September has the highest number of admissions with a mean of 2.88 per day. The winter months are quite low in comparison with mean values from 0.97 to 1.34.

Each pollutant was also plotted against time with the raw plots included as Figure 7 and the LOESS smoothed plots as Figure 8. Looking at Figure 8, we see that $PM_{2.5}$,

PM_{10}, NO_2, SO_2, and CO all demonstrated a modest, yet steady decline in their daily levels. The plot pertaining to ozone shows that the levels of ozone rose and dropped very evenly and consistently over the course of the period. The decrease occurring at the end of the time period was very small and fairly insignificant. As was the case with the daily admission counts, the peaks on the ozone plot are occurring in the warmer seasons while the cooler seasons experienced the lower levels.

The parameter estimates, standard errors and *t*-statistics were calculated using time series and case-crossover methods, and results are provided in Table 2A. The odds ratios, as well as the 95% confidence intervals for model (2.1) are given in Table 3A. The parameter estimates, standard errors and *t*-statistics for model (2.2) are included in Table 2B and the odds ratios and confidence intervals for this analysis are included in Table 3B.

Of all the pollutants, nitrogen dioxide and carbon monoxide show the most significant connections to the admissions with odds ratios and confidence intervals averaging 1.017 and (0.980, 1.056), and 1.026 and (0.979, 1.050) respectively. In other words, the odds ratio of 1.017 means that asthma hospital admissions are 1.017 times more likely with every increase of 11 units (the inter-quartile range) in NO_2.

The three ways of estimation in bidirectional designs, (Breslow, Efron, and Exact) showed similar results for both NO_2 and CO with odds ratios ranging from 1.012 to 1.016 for NO_2, and from 1.001 to 1.004 for CO. The results from the unidirectional designs, however, were not so agreeable. With NO_2, the design comparing a case date to a previous control date produced a higher odds ratio whereas the design comparing a case date to a later control date using CO has a larger value. This inconsistency is observed in

the other pollutants as well with $PM_{2.5}$ and SO_2 producing higher results under the pre-unidirectional design and PM_{10} and ozone producing higher results under the post-unidirectional design.

The odds ratios for the unidirectional design were much lower once the weather variables were accounted for under both NO_2 and CO, and the confidence intervals here were much wider. On the other hand, the values from the bidirectional and time series models hardly changed at all between the two analyses.

Ozone averaged an odds ratio of 0.892 among the different methods of analysis, before being adjusted for weather conditions. The confidence interval for this ratio is (0.853, 0.934). This odds ratio is very low (much smaller than 1) which indicates that there is no association between the number of asthma hospitalizations and the level of ozone in Toronto. The post-unidirectional design gave the highest odds ratio of 1.018 and the pre-unidirectional design had the lowest odds ratio with a value of only 0.848.

Once the model had been adjusted to control the confounding weather conditions, the results changed dramatically. The three bidirectional designs in case-crossover had a combined odds ratio of 0.871 in the initial analysis. The weather adjusted bidirectional design averaged a much lower odds ratio of 0.817. Previously the largest and smallest values for ozone were obtained by post- and pre-unidirectional designs respectively. The time series pre-adjustment showed very little change between the two models.

The odds ratio and confidence interval for SO_2 averaged 0.984 and (0.948, 1.021) which shows that it does not have a strong association with asthma hospitalizations. The values obtained with SO_2 follow similar patterns as those found in the PM_{10} analysis. The pre-unidirectional gave an odds ratio of 1.035 and the post-unidirectional gave one of

0.98, both with confidence intervals of about 0.08 points in width. These odds ratios dropped significantly (0.829 for pre and 0.875 for post) and the confidence intervals grew tremendously in width (from 0.08 points to approximately 25 points) when temperature and humidity were added to the model. Bidirectional averaged 0.965 for the odds ratio under both the adjusted and non-adjusted designs, and time series averaged 0.985 for both designs.

Particulate matter measuring 10 microns (PM_{10}) and 2.5 microns ($PM_{2.5}$) both show a small association with the number of hospital admissions, with $PM_{2.5}$ being the stronger of the two. Analysis of $PM_{2.5}$ produced an averaged odds ratio and confidence interval of 0.984 and (0.954, 1.016), whereas PM_{10} produced an averaged odds ratio of 0.975 and confidence interval of (0.943, 1.008). Each of the models within the three methods (unidirectional, bidirectional, and time series) produced very similar results for $PM_{2.5}$. Both the co-adjustment and the pre-adjustment time series models had values of 0.99. All three bidirectional models (Breslow, Efron, and exact) produced odds ratios of 0.96 and both unidirectional models (pre, post) were evaluated to be 1.00.

The results from the PM_{10} analysis were not as uniform as those from $PM_{2.5}$. The bidirectional and time series designs had similar results among the different models, but the pre-unidirectional value of 0.975 was much lower than the post-unidirectional value of 1.012. Also, the confidence intervals under $PM_{2.5}$ were slightly smaller than those for PM_{10}.

While the odds ratios dropped slightly for PM_{10} and $PM_{2.5}$ (0.01 points on average) under the bidirectional and time series weather adjusted models, both of the unidirectional models generated odds ratio values much lower than those from the

unadjusted models. $PM_{2.5}$ dropped from 1.00 to 0.79 and PM_{10} dropped from an average of 0.99 to that of 0.83. A similar trend was observed in the confidence intervals as well. The bidirectional and time series confidence intervals held essentially the same width with weather accounted for, however, the unidirectional models widened considerably. $PM_{2.5}$ began as (0.97, 1.04) and became (0.04, 15.74) and likewise for PM_{10}.

2.4 Discussion

Most of the studies on the health effects of air pollution that have been done in the past have been performed in the United States of America (Bascom *et al.*, 1996; Schwartz *et al.*, 1993; Neas *et al.*, 1999; Schwartz and Morris, 1995). More and more, studies are being carried out in European cities as well (Castellsaqué *et al.*, 1995; Sunyer *et al.*, 1997; Scarlett *et al.*, 1996; Mackenbach *et al.*, 1993). While a few studies regarding air pollution and our health have been performed in Canada, (Burnett and Krewski (1994), Bates *et al.* (1990) Bates and Sizto (1987)) this thesis looks at the ambient air pollution concentrations and their effects on asthma in children for Toronto, Ontario.

The nitrogen dioxide results for our data agree with the results found by Castellsaqué *et al.* (1995), Devalia *et al.* (1996) and Burnett *et al.* (1999), in that there is indeed a connection between NO_2 and asthma admissions. G. D'Amato *et al.* (1999) also noted the association of NO_2 with one's respiratory system.

Although carbon monoxide shows an association here, other studies have reported that CO does not affect airway function, thus would not contribute to an asthma attack (D'Amato, 1999). On the other hand, CO is highly correlated with other pollutants, mainly NO_2, and trying to study the effects of CO independent of NO_2 is extremely

difficult, (Burnett *et al.*, 1999). This correlation is perhaps the reason why our study showed CO as being associated with asthmatic incidences, when in actuality it is not.

While this association of CO with asthma hospitalizations coincides with that of Nicolai (1999) and Tobías *et al.* (1996), it contradicts the findings of Schwartz *et al.* (1996), which shows a very high association of CO with hospital admissions. This discrepancy may be due to the different groups being studied. Schwartz looked at the hospital admissions of the elderly in Cleveland, Ohio (not children), and had a daily mean of 22, much higher than our 1.54. Also, his study took all respiratory problems into consideration, whereas our study looked solely at asthma incidences.

Nicolai (1999) states that exposure studies have shown that though asthmatics are susceptible to SO_2 exposure, (Devalia *et al.*, 1996), new asthma cases are not caused by SO_2 exposure. Nicolai also states that the effects of SO_2 are intensified with physical exercise.

Ozone is another one of our pollutants whose results do not absolutely concur with past studies. Tobías *et al.* (1999) showed that with increases of O_3, there were increases in the number of asthma related admissions. However, Barcelona, Spain was the focus of Tobías' study and the O_3/asthma association there had not yet been reported. The results presented in this thesis show that O_3 levels do not contribute to asthma incidences, since O_3 has odds ratios very close to 1. Weather conditions were more severe (higher temperatures and higher percentages of humidity), and the pollutant levels, as well as the admission counts were higher in Barcelona than in Toronto during the period in question. These differences may account for the conflicting results with this paper.

As previously explained, ozone is the result of the reaction occurring when nitrogen dioxide and vehicular exhaust are simultaneously exposed to sunlight. Thus the pattern observed in the plot of ozone against time is not surprising, since the percentage of sunny days is much higher in the summer than in the winter, it is logical that more ozone would be created during the summer months.

Since children spend less time outdoors in the winter than in the summer, they are subjected to less pollutant exposure during the winter months. This explains the recurring, annual pattern of the admission plots. Some of the pollutants tend to reach lower daily values during the winter months, which may also reduce the risk of asthma attacks resulting in hospitalization.

In summary, over the course of the study period, the admission counts and most of the pollutant levels were decreasing in value. The only pollutants that were found to be significantly associated with asthma hospitalizations were nitrogen dioxide and carbon monoxide. Adjusting for the confounding weather variables actually decreased the odds ratios for all of the pollutants.

The results of the analyses, concerning the relationship between air pollution and asthma hospitalizations, are somewhat puzzling since the different methods of analyses did not give consistent answers. While analyzing certain pollutants, the time series approach seems to give a higher odds ratio, showing more of an association with asthma hospitalizations. On the other hand, some of the pollutants produced a higher odds ratio under the unidirectional case-crossover design, and others under the bidirectional case-crossover design.

Due to this inconsistency, it is hard to draw conclusions as to the relationship between air pollution and asthma hospitalizations. In order to see which of the methods produce the most accurate result, a simulation was performed to evaluate each of the time series and case-crossover designs, using the pollutant $PM_{2.5}$ as an example. The results and the discussion of those results are presented in Chapter 3.

Table 1. Statistical summary of daily hospital admissions for children ages 6-12, daily weather conditions, and daily concentrations of particulate and gaseous pollutants, Toronto 1981 to 1993.

Hospital Admissions Summary	Females	Males	Males and Females
Mean	0.57	0.97	1.54
Standard Deviation	0.8063	1.1414	1.5145
Maximum Value for Period	5	11	11
Minimum Value for Period	0	0	0

Weather Conditions Summary	Maximum Daily Temperature	Minimum Daily Temperature	Relative Humidity
Mean	12.68	2.63	73.51
Standard Deviation	11.28	9.91	11.30
75^{th} Percentile	22.5	10.8	81
25^{th} Percentile	3.2	-3.7	66
Inter-quartile Range	19.3	14.5	15
Maximum Value for Period	37.60	24.30	99
Minimum Value for Period	-19.80	-31.30	35
Zero Counts	19	35	0

Pollutant Value Summary	PM_{10}	$PM_{2.5}$	CO	NO_2	SO_2	O_3
Mean	29.77	18.01	1.18	25.03	5.34	30.20
Standard Deviation	13.08	8.61	0.50	9.07	5.86	17.21
75^{th} Percentile	35.32	21.74	1.40	30	8	38
25^{th} Percentile	21.01	12.37	0.90	19	1	18
Inter-quartile Range	14.31	9.36	0.50	11	7	20
Maximum Value for Period	99.58	89.59	6.10	82	56	98
Minimum Value for Period	0.08	1.22	0	3	0	0
Zero Counts	0	0	5	0	890	2

Abbreviations: PM_{10}, particulate matter smaller than 10 microns in diameter; $PM_{2.5}$, particulate matter smaller than 2.5 microns in diameter; CO, carbon monoxide; NO_2, nitrogen dioxide; SO_2, sulfur dioxide; O_3, ozone.

Table 2A. Estimates, standard errors and *t*-statistics for each pollutant.

Abbreviations: PM_{10}, particulate matter smaller than 10 microns in diameter; $PM_{2.5}$, particulate matter smaller than 2.5 microns in diameter; CO, carbon monoxide; NO_2, nitrogen dioxide; SO_2, sulfur dioxide; O_3, ozone.

$PM_{2.5}$		$\hat{\beta}_1$	*Standard Error*	*t*
Unidirectional	Pre	0.0004	0.0019	0.2103
	Post	0.0004	0.0019	0.1929
Bidirectional	Breslow	-0.0034	0.0017	-2.0697
	Efron	-0.0040	0.0017	-2.3818
	Exact	-0.0043	0.0019	-2.3277
Time Series	Co-Adjustment 93	-0.0004	0.0014	-0.2549
	Pre-Adjustment 93	-0.0006	0.0015	-0.4134

PM_{10}		$\hat{\beta}_1$	*Standard Error*	*t*
Unidirectional	Pre	-0.0015	0.0013	-1.1393
	Post	0.0009	0.0014	0.6377
Bidirectional	Breslow	-0.0025	0.0011	-2.1964
	Efron	-0.0030	0.0012	-2.5658
	Exact	-0.0031	0.0013	-2.4619
Time Series	Co-Adjustment 186	-0.0019	0.0010	-1.8808
	Pre-Adjustment 186	-0.0014	0.0010	-1.3407

NO_2		$\hat{\beta}_1$	*Standard Error*	*t*
Unidirectional	Pre	0.0028	0.0020	1.3679
	Post	0.0008	0.0020	0.3908
Bidirectional	Breslow	0.0011	0.0017	0.6611
	Efron	0.0014	0.0017	0.8545
	Exact	0.0014	0.0019	0.7408
Time Series	Co-Adjustment 186	0.0019	0.0015	1.3231
	Pre-Adjustment 186	0.0016	0.0015	1.0541

SO_2		$\hat{\beta}_1$	*Standard Error*	*t*
Unidirectional	Pre	0.0049	0.0031	1.5859
	Post	-0.0029	0.0030	-0.9502
Bidirectional	Breslow	-0.0045	0.0026	-1.7254
	Efron	-0.0051	0.0026	-1.9365
	Exact	-0.0056	0.0029	-1.9260
Time Series	Co-Adjustment 186	-0.0020	0.0022	-0.9056
	Pre-Adjustment 186	-0.0013	0.0023	-0.5630

CO		$\hat{\beta}_1$	*Standard Error*	*t*
Unidirectional	Pre	0.0566	0.0420	1.3457
	Post	0.0386	0.0413	0.9341
Bidirectional	Breslow	0.0021	0.0344	0.0596
	Efron	0.0073	0.0346	0.2098
	Exact	0.0026	0.0387	0.0671
Time Series	Co-Adjustment 93	0.0383	0.0272	1.4063
	Pre-Adjustment 93	0.0421	0.0324	1.3002

O_3		$\hat{\beta}_1$	*Standard Error*	*t*
Unidirectional	Pre	-0.0082	0.0014	-5.7775
	Post	0.0009	0.0014	0.6201
Bidirectional	Breslow	-0.0061	0.0011	-5.7035
	Efron	-0.0070	0.0011	-6.5901
	Exact	-0.0076	0.0012	-6.3747
Time Series	Co-Adjustment 186	-0.0068	0.0008	-8.5836
	Pre-Adjustment 186	-0.0054	0.0011	-5.1504

Table 2B. Estimates, standard errors and *t*-statistics for each pollutant under weather adjusted model.

Abbreviations: PM_{10}, particulate matter smaller than 10 microns in diameter; $PM_{2.5}$, particulate matter smaller than 2.5 microns in diameter; CO, carbon monoxide; NO_2, nitrogen dioxide; SO_2, sulfur dioxide; O_3, ozone.

$PM_{2.5}$		$\hat{\beta}_1$	*Standard Error*	*t*
Unidirectional	Pre	-0.0253	0.1654	-0.1528
	Post	-0.0242	0.1603	-0.1512
Bidirectional	Breslow	-0.0049	0.0018	-2.6664
	Efron	-0.0058	0.0019	-3.1144
	Exact	-0.0061	0.0021	-2.9923
Time Series	Co-Adjustment 93	-0.0018	0.0015	-1.2026
	Pre-Adjustment 93	-0.0004	0.0016	-0.2650

PM_{10}		$\hat{\beta}_1$	*Standard Error*	*t*
Unidirectional	Pre	-0.0123	0.1098	-0.1121
	Post	-0.0136	0.1061	-0.1281
Bidirectional	Breslow	-0.0035	0.0013	-2.7293
	Efron	-0.0042	0.0013	-3.2124
	Exact	-0.0044	0.0014	-3.0472
Time Series	Co-Adjustment 186	-0.0028	0.0011	-2.4904
	Pre-Adjustment 186	-0.0018	0.0011	-1.6001

NO_2		$\hat{\beta}_1$	*Standard Error*	*t*
Unidirectional	Pre	-0.0224	0.1546	-0.1448
	Post	-0.0186	0.1596	-0.1167
Bidirectional	Breslow	0.0011	0.0018	0.6431
	Efron	0.0015	0.0018	0.8314
	Exact	0.0014	0.0020	0.7173
Time Series	Co-Adjustment 186	0.0017	0.0015	1.0889
	Pre-Adjustment 186	0.0014	0.0016	0.9090

SO_2		$\hat{\beta}_1$	*Standard Error*	*t*
Unidirectional	Pre	-0.0268	0.2518	-0.1067
	Post	-0.0191	0.2421	-0.0789
Bidirectional	Breslow	-0.0046	0.0027	-1.6760
	Efron	-0.0051	0.0027	-1.8640
	Exact	-0.0057	0.0030	-1.866
Time Series	Co-Adjustment 186	-0.0023	0.0022	-0.9904
	Pre-Adjustment 186	-0.0013	0.0024	-0.5479

CO		$\hat{\beta}_1$	*Standard Error*	*t*
Unidirectional	Pre	-0.4087	3.0856	-0.1325
	Post	-0.4273	3.2252	-0.1326
Bidirectional	Breslow	-0.0079	0.0363	-0.2178
	Efron	-0.0049	0.0365	-0.1341
	Exact	-0.0100	0.0408	-0.2449
Time Series	Co-Adjustment 93	0.0215	0.0288	0.7456
	Pre-Adjustment 93	0.0426	0.0322	1.3239

O_3		$\hat{\beta}_1$	*Standard Error*	*t*
Unidirectional	Pre	0.0004	0.1120	0.0039
	Post	-0.0081	0.1074	-0.0756
Bidirectional	Breslow	-0.0089	0.0013	-6.8755
	Efron	-0.0104	0.0013	-8.0428
	Exact	-0.0110	0.0014	-7.6553
Time Series	Co-Adjustment 186	-0.0088	0.0011	-8.2290
	Pre-Adjustment 186	-0.0058	0.0011	-5.1968

Table 3A. Odds ratios and confidence intervals for each pollutant.

Abbreviations: PM_{10}, particulate matter smaller than 10 microns in diameter; $PM_{2.5}$, particulate matter smaller than 2.5 microns in diameter; CO, carbon monoxide; NO_2, nitrogen dioxide; SO_2, sulfur dioxide; O_3, ozone.

$PM_{2.5}$		*Odds Ratio*	*Confidence Interval*
Unidirectional	Pre	1.0038	0.9689, 1.0400
	Post	1.0035	0.9687, 1.0395
Bidirectional	Breslow	0.9684	0.9393, 0.9983
	Efron	0.9631	0.9338, 0.9934
	Exact	0.9603	0.9280, 0.9936
Time Series	Co-Adjustment 93	0.9967	0.9716, 1.0224
	Pre-Adjustment 93	0.9941	0.9665, 1.0225

PM_{10}		*Odds Ratio*	*Confidence Interval*
Unidirectional	Pre	0.9784	0.9422, 1.0159
	Post	1.0124	0.9748, 1.0515
Bidirectional	Breslow	0.9648	0.9343, 0.9961
	Efron	0.9585	0.9279, 0.9900
	Exact	0.9560	0.9224, 0.9909
Time Series	Co-Adjustment 186	0.9735	0.9466, 1.0011
	Pre-Adjustment 186	0.9803	0.9521, 1.0093

NO_2		*Odds Ratio*	*Confidence Interval*
Unidirectional	Pre	1.0310	0.9869, 1.0771
	Post	1.0086	0.9663, 1.0526
Bidirectional	Breslow	1.0122	0.9765, 1.0492
	Efron	1.0159	0.9798, 1.0535
	Exact	1.0154	0.9752, 1.0571
Time Series	Co-Adjustment 186	1.0215	0.9898, 1.0542
	Pre-Adjustment 186	1.0177	0.9850, 1.0515

SO_2		*Odds Ratio*	*Confidence Interval*
Unidirectional	Pre	1.0350	0.9919, 1.0798
	Post	0.9801	0.9404, 1.0216
Bidirectional	Breslow	0.9691	0.9352, 1.0043
	Efron	0.9651	0.9310, 1.0004
	Exact	0.9618	0.9244, 1.0007
Time Series	Co-Adjustment 186	0.9859	0.9562, 1.0166
	Pre-Adjustment 186	0.9909	0.9600, 1.0229

CO		*Odds Ratio*	*Confidence Interval*
Unidirectional	Pre	1.0287	0.9872, 1.0720
	Post	1.1095	0.9790, 1.0616
Bidirectional	Breslow	1.0010	0.9679, 1.0353
	Efron	1.0036	0.9702, 1.0383
	Exact	1.0013	0.9641, 1.0400
Time Series	Co-Adjustment 93	1.0193	0.9925, 1.0469
	Pre-Adjustment 93	1.0213	0.9894, 1.0542

O_3		*Odds Ratio*	*Confidence Interval*
Unidirectional	Pre	0.8484	0.8023, 0.8970
	Post	1.0176	0.9630, 1.0753
Bidirectional	Breslow	0.8846	0.8481, 0.9227
	Efron	0.8687	0.8330, 0.9058
	Exact	0.8586	0.8193, 0.8998
Time Series	Co-Adjustment 186	0.8702	0.8464, 0.9005
	Pre-Adjustment 186	0.8971	0.8608, 0.9349

Table 3B. Odds ratios and confidence intervals for each pollutant under weather adjusted model.

Abbreviations: PM_{10}, particulate matter smaller than 10 microns in diameter; $PM_{2.5}$, particulate matter smaller than 2.5 microns in diameter; CO, carbon monoxide; NO_2, nitrogen dioxide; SO_2, sulfur dioxide; O_3, ozone.

$PM_{2.5}$		*Odds Ratio*	*Confidence Interval*
Unidirectional	Pre	0.7893	0.0380, 16.4085
	Post	0.7970	0.0421, 15.0863
Bidirectional	Breslow	0.9554	0.9239, 0.9880
	Efron	0.9473	0.9156, 0.9813
	Exact	0.9443	0.9095, 0.9804
Time Series	Co-Adjustment 93	0.9831	0.9562, 1.0108
	Pre-Adjustment 93	0.9962	0.9682, 1.0249

PM_{10}		*Odds Ratio*	*Confidence Interval*
Unidirectional	Pre	0.8386	0.0386, 18.2396
	Post	0.8233	0.0420, 16.1276
Bidirectional	Breslow	0.9512	0.9176, 0.9860
	Efron	0.9421	0.9084, 0.9770
	Exact	0.9395	0.9026, 0.9780
Time Series	Co-Adjustment 186	0.9613	0.9319, 0.9916
	Pre-Adjustment 186	0.9753	0.9458, 1.0057

NO_2		*Odds Ratio*	*Confidence Interval*
Unidirectional	Pre	0.7818	0.0279, 21.8994
	Post	0.8148	0.0261, 25.4368
Bidirectional	Breslow	1.0125	0.9749, 1.0514
	Efron	1.0163	0.9783, 1.0557
	Exact	1.0156	0.9735, 1.0596
Time Series	Co-Adjustment 186	1.0183	0.9856, 1.0520
	Pre-Adjustment 186	1.0159	0.9819, 1.0511

SO_2		*Odds Ratio*	*Confidence Interval*
Unidirectional	Pre	0.8285	0.0262, 26.2290
	Post	0.8749	0.0352, 24.2467
Bidirectional	Breslow	0.9686	0.9332, 1.0054
	Efron	0.9648	0.9292, 1.0018
	Exact	0.9613	0.9222, 1.0020
Time Series	Co-Adjustment 186	0.9840	0.9530, 1.0159
	Pre-Adjustment 186	0.9908	0.9584, 1.0242

CO		*Odds Ratio*	*Confidence Interval*
Unidirectional	Pre	0.8152	0.0396, 16.7685
	Post	0.8076	0.0342, 19.0496
Bidirectional	Breslow	0.9961	0.9613, 1.0321
	Efron	0.9976	0.9625, 1.0339
	Exact	0.9950	0.9560, 1.0356
Time Series	Co-Adjustment 93	1.0108	0.9827, 1.0398
	Pre-Adjustment 93	1.0215	0.9898, 1.0543

O_3		*Odds Ratio*	*Confidence Interval*
Unidirectional	Pre	1.0887	0.0125, 81.2333
	Post	0.8502	0.0126, 57.1884
Bidirectional	Breslow	0.8374	0.7961, 0.8809
	Efron	0.8116	0.7713, 0.8539
	Exact	0.8028	0.7589, 0.8492
Time Series	Co-Adjustment 186	0.8384	0.8039, 0.8743
	Pre-Adjustment 186	0.8913	0.8534, 0.9308

Figure 1. Comparison of various autocorrelations function (ACF) plots.

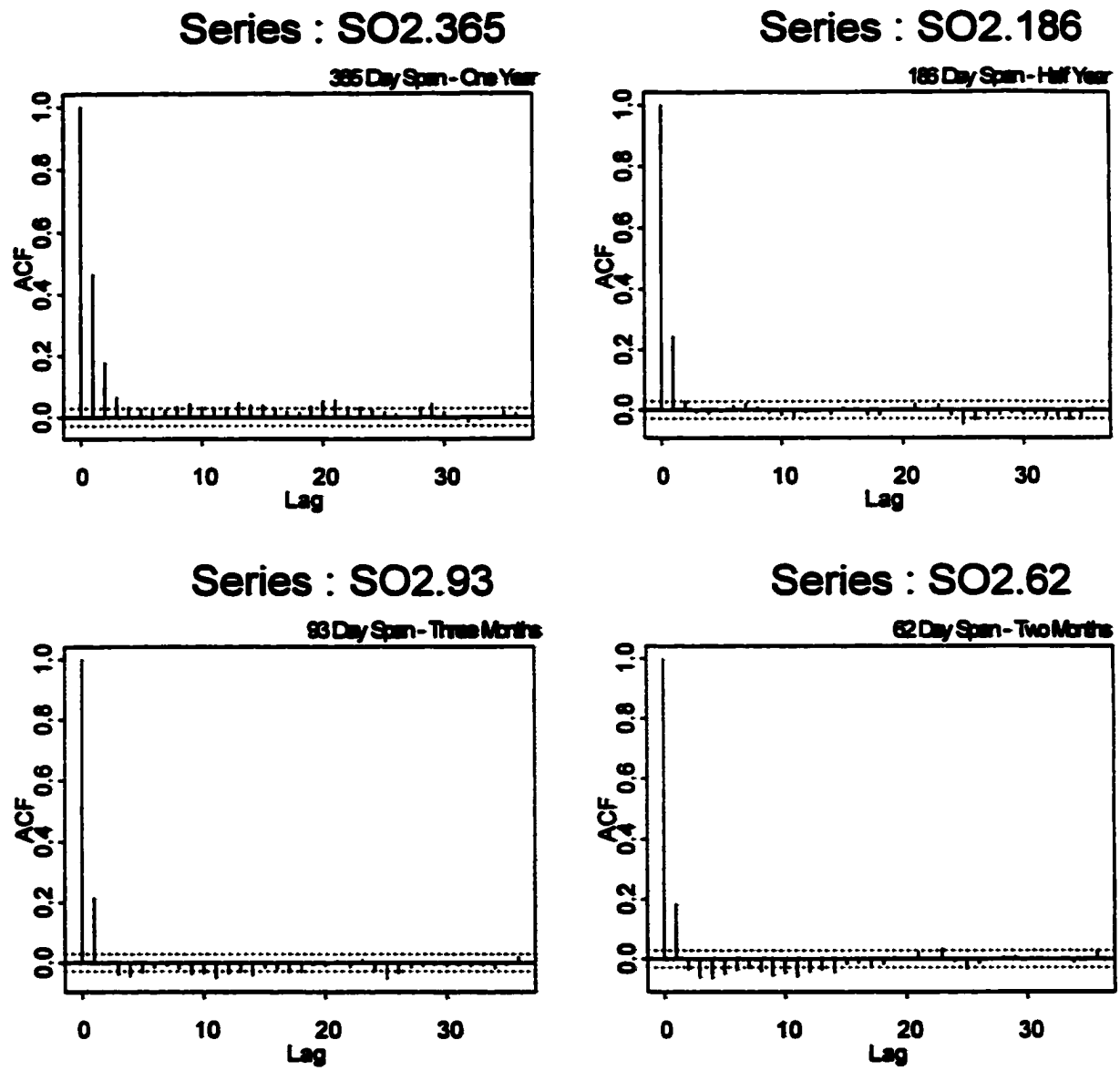


Figure 2. Autocorrelation (ACF) plots for each pollutant illustrating the most appropriate span.

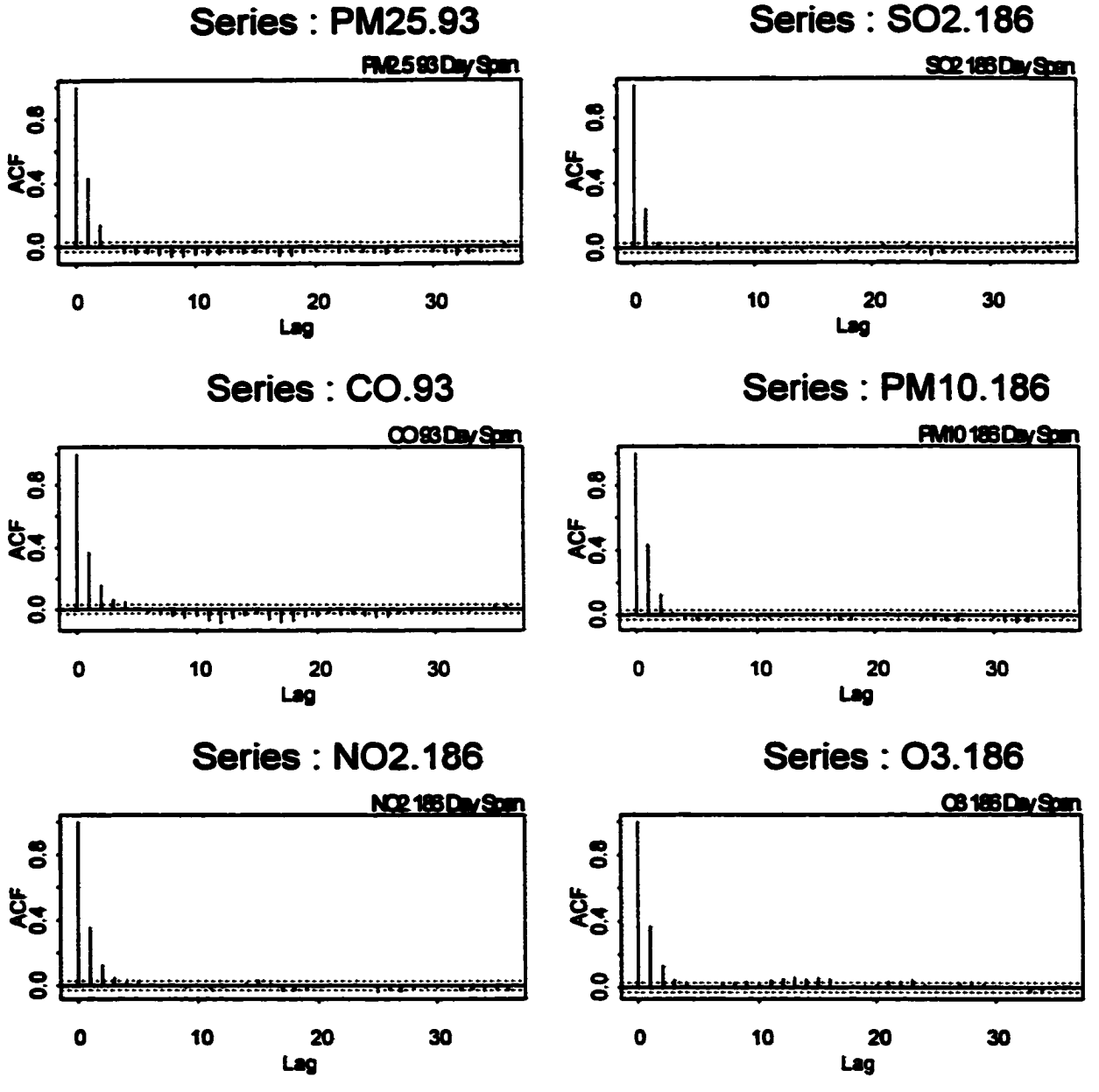
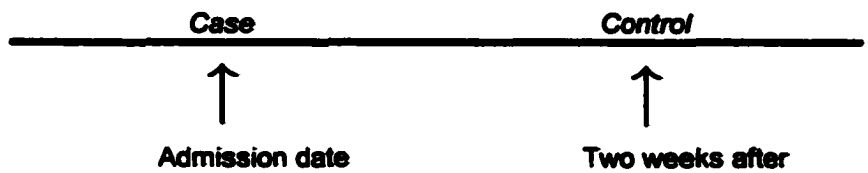


Figure 3 - Case-crossover diagram

Unidirectional (Pre)

Unidirectional (Post)



Bidirectional Control

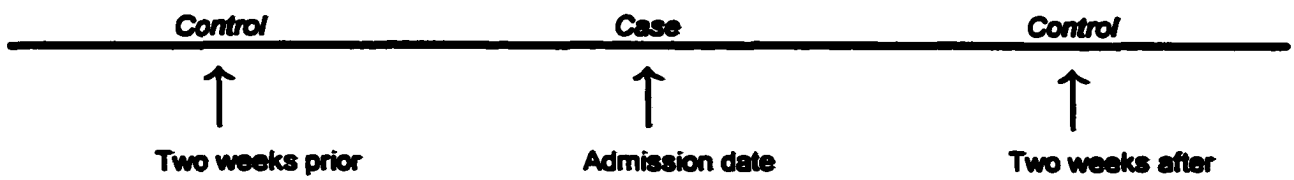


Figure 4. Smoothed time series plots of daily admissions for females, males and the two combined against time in days.

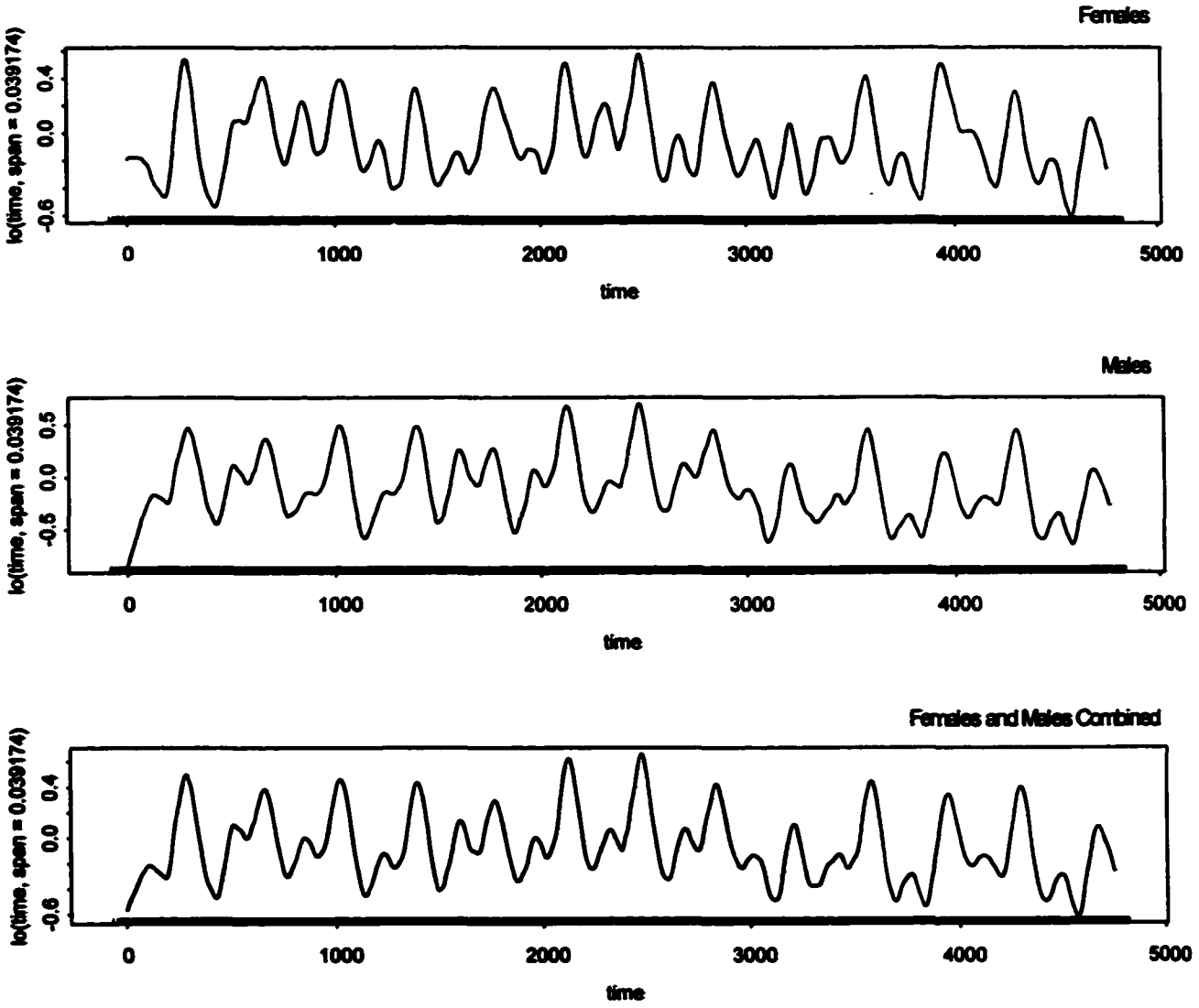
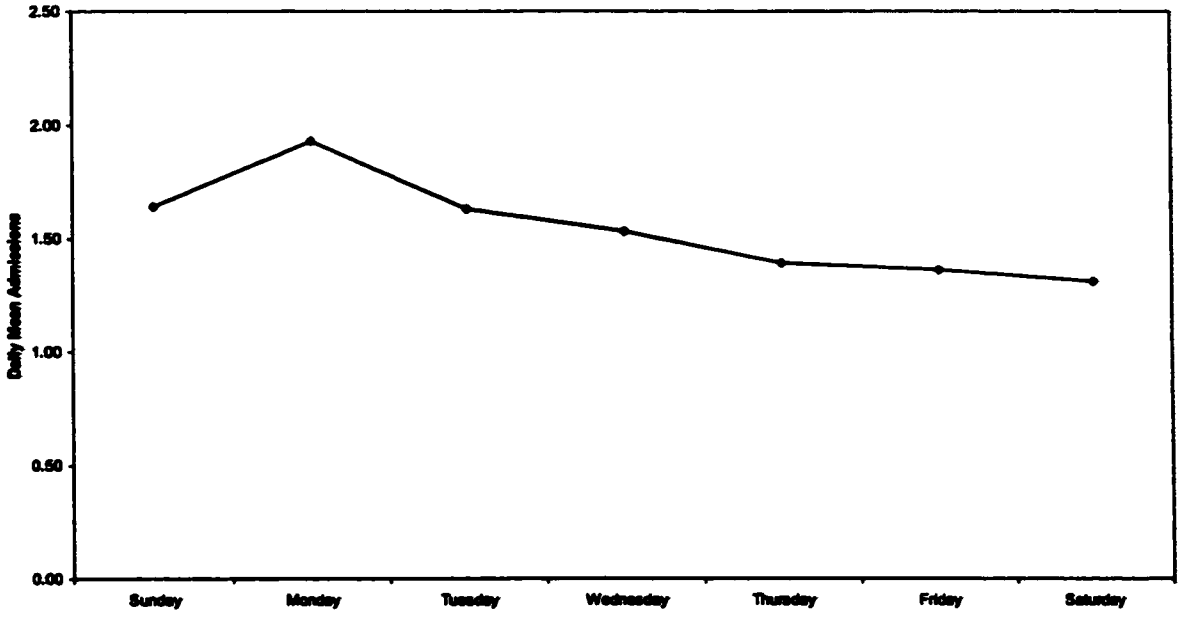


Figure 5. Daily mean of asthma admissions to the hospital by day of the week and by month in Toronto, Ontario for the period of 1981 thru 1993.



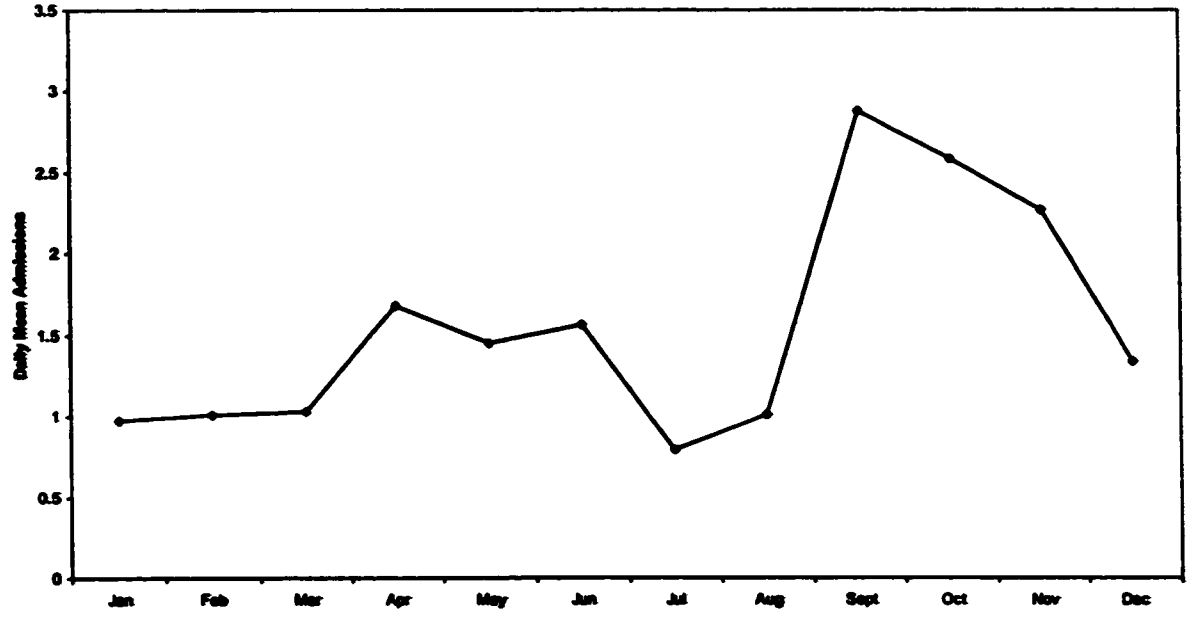


Figure 6. Unsmoothed (raw) plots of daily hospital admissions for females, males and the two combined against time in days.

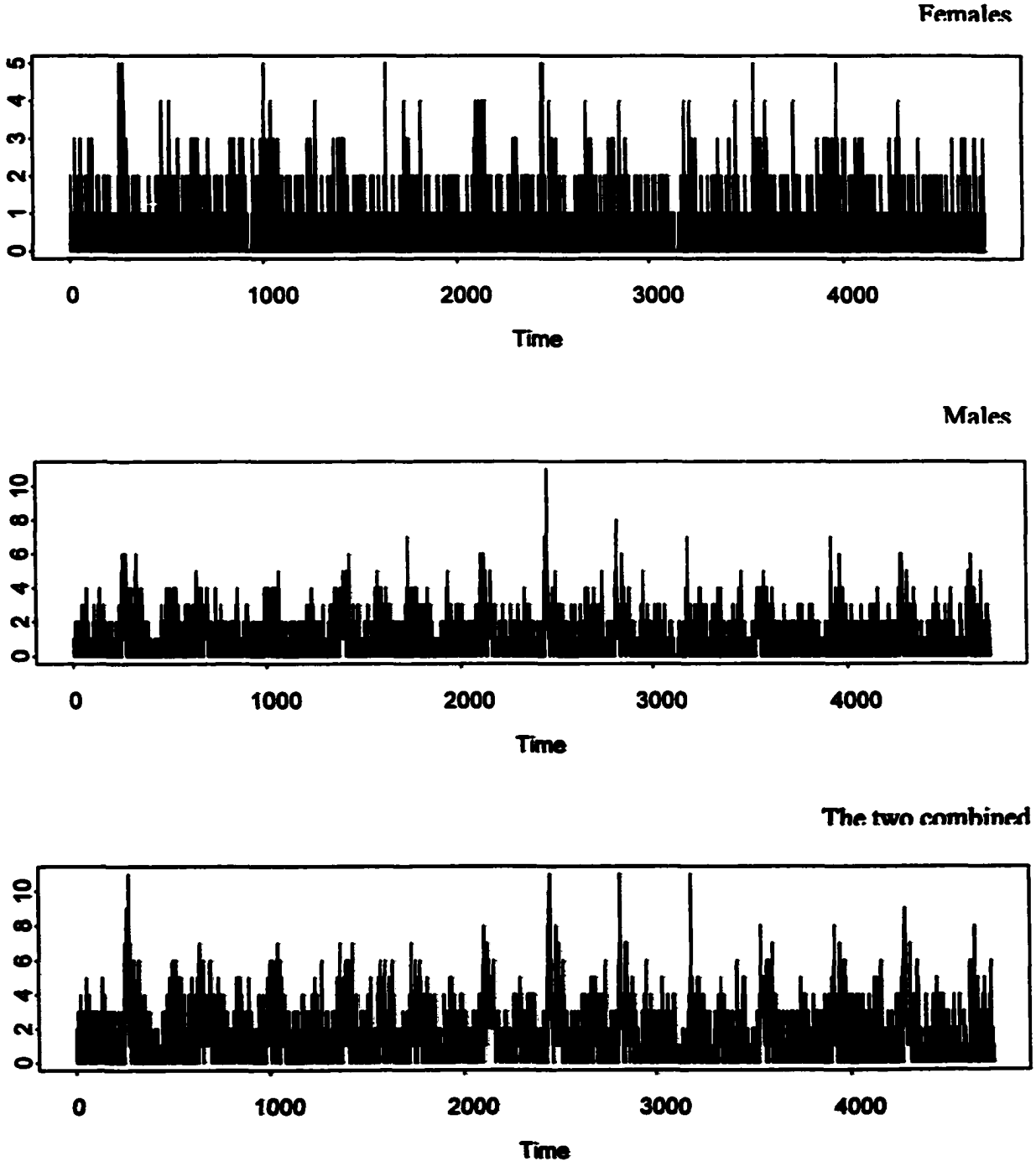


Figure 7. Unsmoothed (raw) plots of each pollutant against time in days.

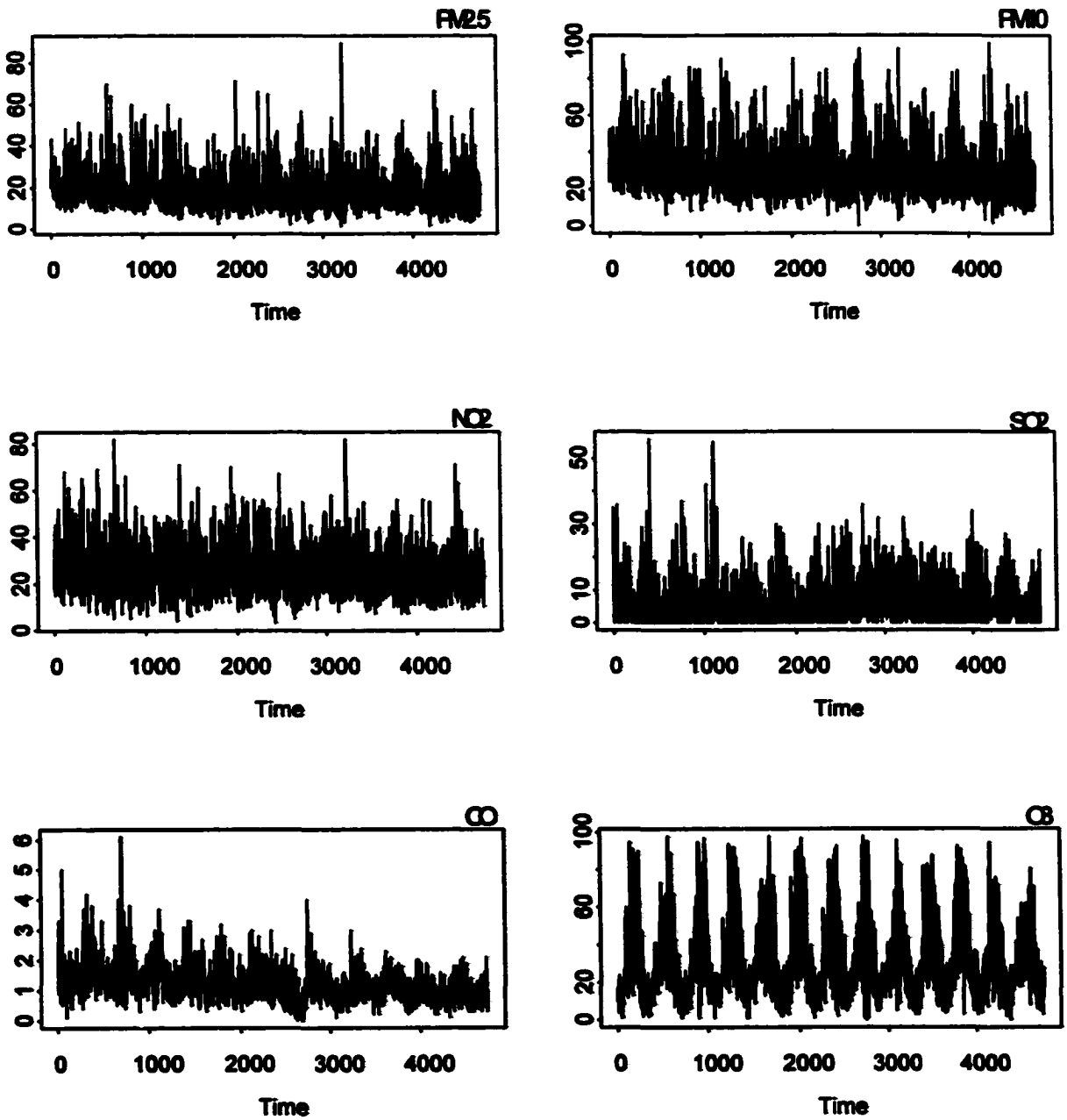
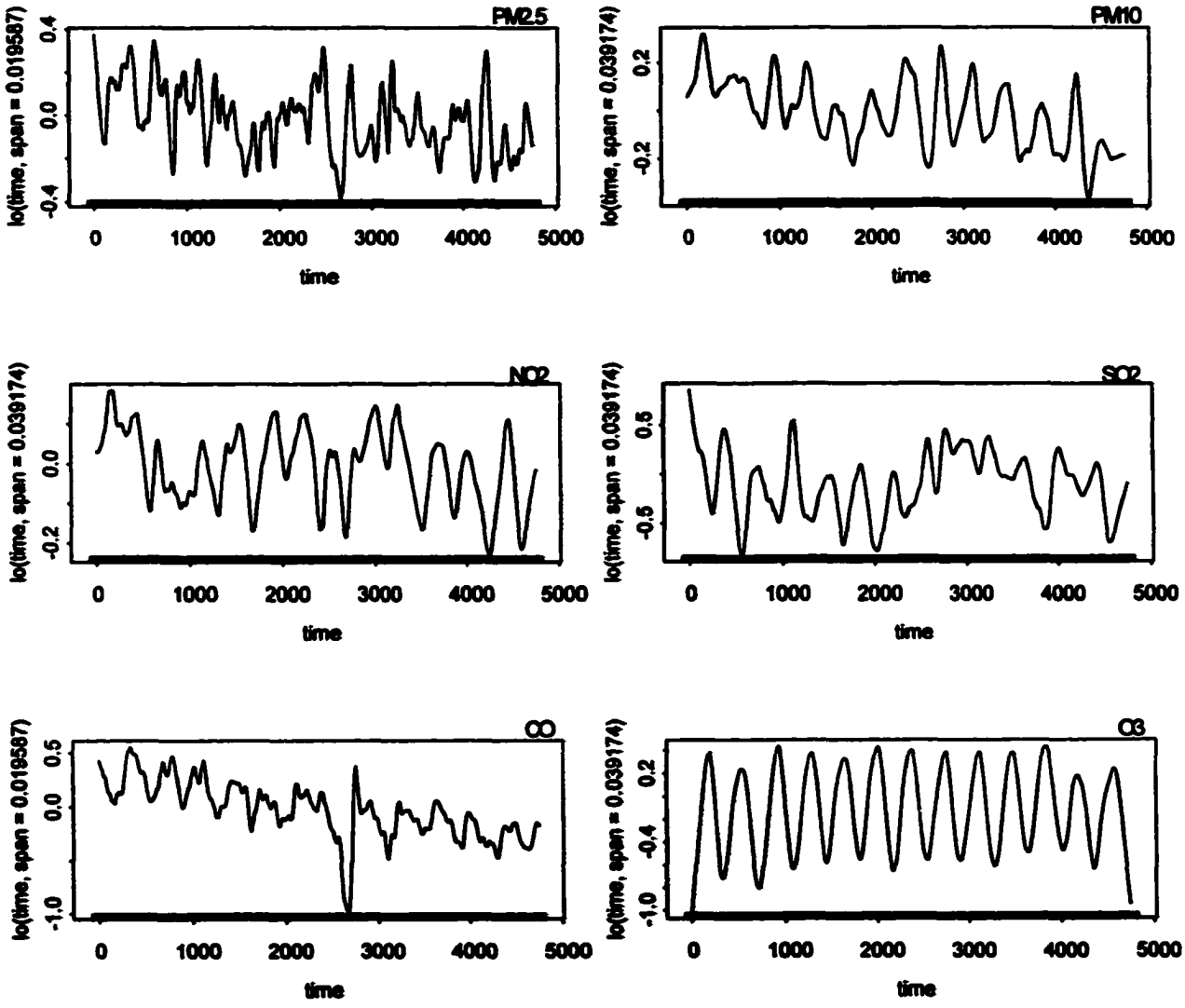


Figure 8. Smoothed plots of each pollutant against time in days.



Chapter 3

Simulation Study

As mentioned in Chapter 2, time series and case-crossover are two of the more popular methods of analysis since many variations are available to best suit the data at hand. The problem with having such a variety of analyses is that it is hard to choose the best method. Comparing the estimates to the true parameter value is a way of evaluation, but the problem is that the true value is not usually known. Hence, we resort to computer simulations.

Simulations allow us to choose the models and parameters ahead of time and data can be generated accordingly. We can compare the estimates to the parameter values once the simulation has been performed.

The computer software S-Plus was used to obtain all of the results in this thesis. All of the methods used to perform the simulation will be discussed in section 3.1. Also, the results of the simulation and a discussion about them are included in sections 3.2 and 3.3.

3.1 Methods

Simulations are wonderful tools that allow us to control certain aspects of a model in order to investigate others. Simulations can be used to better understand the results obtained from an analysis of raw data. Due to the similarity of simulating each of the air

pollutants, a simulation of $PM_{2.5}$ was the only one performed here. $PM_{2.5}$ was chosen because small airborne particles are known suspects in the cause of asthmatic attacks. The second chapter of this thesis found a small association between $PM_{2.5}$ and hospital admissions. Also, there have been studies done in the past which have positively shown $PM_{2.5}$ to be a cause of respiratory related admissions (see for example Schwartz and Dockery, 1992).

Simulations can be very time consuming, as they require many iterations of the experiment and the application of many computer routines that are sometimes fairly lengthy. Even the fastest computers have a difficult time processing some of these programs quickly. It is because of this and the similarity of simulating females, males or the two combined, that the $PM_{2.5}$ simulation was done using only the male admission counts. Also, it is more common for young boys to be diagnosed with asthma than it is for girls.

To keep things simple, temperature and humidity were not included in the simulation. Filters were used in the simulation to account for temporal trends.

A Poisson model was used to generate the asthma hospitalizations. The number of admissions on the t^{th} day, y_t, has mean

$$E(y_t) = e^{(\beta_0 + \beta_1 x_t)} * D_t * S_t \qquad \text{Model I}$$

where D_t refers to the day of the week filter and S_t refers to the Shumway filter. This model uses the observed $PM_{2.5}$ values in x_t to generate the admissions. The time period for the data spans 4748 days (from January 1, 1981 to December 31, 1993).

The admission weights, D_t, for each day of the week as were used in Burnett and Krewski (1994), (1.16(Mondays), 1.07, 1.00, 1.01, 1.01, 0.84, 0.89(Sundays)), were also

used in this simulation. This information is very important since each day does not have the same probability of recording a case. Most people spend more time outdoors on the weekends rather than during the week so it is logical that there is a higher probability that a case will occur just after the weekend (Monday, Tuesday).

Shumway *et al.* (1983) suggest using weights as a filter for temporal trends. The nineteen-day symmetric linear filter on our admissions was

$$S_t = \sum_{k=-9}^{9} \psi_k \left(\frac{u_{t-k}}{\bar{u}} \right)$$

where ψ is given by (0.0874, 0.0857, 0.0807, 0.0729, 0.0629, 0.0518, 0.0404, 0.0329, 0.0200, 0.0123). Here, u_t denotes the number of asthma admissions for boys aged six to twelve years of age in Toronto on the t^{th} day, and $\bar{u}$ is the average number of admissions per day over the entire period.

The simulation uses the actual mean and inter-quartile range, 18.01 and 9.36 $\mu g / m^3$ respectively, of $PM_{2.5}$. The odds ratio was taken to be 1.05, which means that for every increase of 9.36 $\mu g / m^3$ of $PM_{2.5}$, there is an increase in asthma hospitalizations by five percent. The odds ratio is equal to $\exp(\beta_1)$ for every one unit of increase in x. But here the odds ratio is defined as an increase in admissions for every increase in x by the inter-quartile range. Thus the odds ratio is equal to $\exp\left(\frac{\beta_1}{IQR}\right)$, which means that

$$\beta_1 = \ln(1.05)/9.36 = 0.00525.$$

The mean number of admissions per day for boys aged six to twelve is 0.97. If D_t and S_t are assumed to be 1 or very close to 1, we can substitute 0.97 for $E(y_t)$, 18.01 for x_t and 0.00525 for β_1 in Model I, yielding –0.12 for β_0.

$$\beta_0 = \ln(0.97) - \beta_1(18.01) = -0.12$$

Another model was investigated, but this time the generating pollutant has been lagged three days.

$$E(y_t) = e^{(\beta_0 + \beta_1 z_t)} * D_t * S_t \qquad \text{Model II}$$

Every element remains the same as in Model I, but z_t is now the lagged $PM_{2.5}$. The three-day lag is a weighted average of today, yesterday and the day before yesterday with a weight ratio of 1:0.5:0.25. Once these weights were normalized so that they sum to 1, we obtained the following,

$$z_t = 0.571x_t + 0.286x_{t-1} + 0.143x_{t-2}.$$

In this equation, x_t is the pollutant value recorded on the day of incidence, and x_{t-1}, x_{t-2} are the pollutant values 1 and 2 days prior.

The series of y_t in Model I is analyzed using one and three-day averages of $PM_{2.5}$. The one-day average looks solely at today's pollutant and admission count values. The three-day average looks at the average of pollutant levels of today, yesterday and the day before yesterday. It is important to notice that the simulation is generating the admission data, but running the analysis with the observed $PM_{2.5}$ values. The series of y_t in Model II is analyzed using one-day, three-day and three-day lagged averages of $PM_{2.5}$.

The time series technique was applied with two different approaches: pre-adjustment and co-adjustment. In co-adjustment the seasonal trends are being removed at

the same time that the pollutant is being regressed on to the admissions in a generalized additive model (GAM). Alternatively, in pre-adjustment the admissions and pollution time series both have the seasonal trends removed separately. Once the trends have been removed the two series are linked together, and then analyzed in a GAM.

The vector of daily admission counts that was generated under Model I was used as the dependent variable y in the regression of the following GAM.

$$y = lo(time, span = a) + x + D_t$$

In this function, x is the pollutant, D_t is the weight corresponding to the day of the week, and *lo(time,span=a)* is a LOESS smoothing function. In the LOESS function, *time* is a vector, being used as an index for the cases, and *span=a* is the window of the smoothing function. The span used in this simulation was 93 days. This span was chosen based on the autocorrelation function technique described in Chapter 2. The estimates of β_1, standard errors, confidence intervals and the odds ratios were easily obtained by calling the *gam* routine in S-Plus.

We also analyzed the same data sets using the case-crossover design. Unidirectional refers to the comparison of the $PM_{2.5}$ level today with the level on a previous date (pre) or a subsequent date (post). Bidirectional refers to the simultaneous comparison of today's $PM_{2.5}$ level with the levels on a date before and after today. Each of the case-crossover data sets were separately analyzed using conditional logistic regression matched on day and weighted by that day's outcome case count.

Both of these procedures consider today's incidence to be a case and the comparison dates are controls. Every control date is two weeks (14 days) away from the case date. If the time interval between the case and the control is short, the problem of

possible weather or seasonal differences over the interval is taken care of. On the other hand, if the interval is too short, there may be some autocorrelation between the exposures. Taking all of this into consideration, an interval of two weeks was chosen. Diagrams of the unidirectional and bidirectional designs are presented in Figure 3.

This experiment was repeated five hundred times and an average was taken of each of the estimates of β_1 and of the odds ratios. The standard deviation was obtained by taking the square root of the mean of the variances. These results are provided at the end of this chapter.

3.2 Results

Means of β_1's, square root of mean variances and mean odds ratios are given as Table 4. The mean parameter estimates and mean odds ratios obtained from the analyses of the three-day lag generation are included as Table 5.

In Table 4 we see that the time series, co-adjustment design yielded parameter estimates for one-day (0.00395) and three-day (0.00304) averages that are quite lower than the 'true' value of 0.00525. One-day and three-day odds ratios are 1.038 and 1.029 respectively. The results from the analysis under the other time series technique, pre-adjustment, are a bit higher than that of co-adjustment. The one-day and three-day parameter estimates are 0.00447 and 0.00508 respectively. The odds ratio obtained here is 1.049 for both the one and three-day averages.

In the case-crossover analysis, the unidirectional design comparing today's case with the control two weeks prior resulted in very low parameter estimates (0.00271 and 0.00101) for the one-day and three-day averages. The odds ratios are also very low in comparison to the true parameter 1.05, at values of 1.026 (one-day) and 1.009 (three-

day). On the other hand, the unidirectional design comparing today's case to a control two weeks later, yielded values higher than the true parameter and odds ratio, 0.00525 and 1.05. The one-day analysis produced an estimate average of 0.00591 whereas the three-day analysis produced an estimate average of 0.00721. The odds ratios for these two analyses are 1.057 (one-day) and 1.059 (three-day).

Breslow's estimation in the bidirectional case-crossover resulted in estimates and odds ratio values lower than the 'true' values. The mean parameter estimates in Table 4 are 0.00428 (one-day) and 0.00408 (three-day) and the odds ratios are 1.041 (one-day) and 1.039 (three-day). The parameter estimates resulting from the Efron method of estimation are higher than the Breslow estimates but are still fairly low (0.00503 and 0.00471 for one and three-day averages). The odds ratios for the same two averages are 1.048 and 1.045.

Under the exact method of estimation, the parameter estimates and odds ratios are the most accurate of all the bidirectional methods. The estimates for one and three-day analyses under this design are 0.00517 and 0.00491 respectively. The odds ratios here are the closest to the expected odds ratio of 1.05 with values of 1.0496 (one-day) and 1.047 (three-day).

Table 5 shows three types of analyses under each method (one-day, three-day and three-day lagged averages). The average (or mean) estimates for the time series co-adjustment design are 0.00254, 0.00318, and 0.00334 for one-day, three-day and three-day lagged analyses respectively. The odds ratios are 1.024, 1.030, and 1.032.

The results from the time series pre-adjustment method of analysis exhibited a similar trend to that of co-adjustment. Moving from one to three to three-day lagged

averages the mean estimates increased (0.00289, 0.00405, and 0.00410). The odds ratios increased from 1.027 to 1.039 from the one-day to the three-day analyses, but remained at 1.039 for the three-day lag analysis.

The unidirectional pre method severely underestimated the parameter with values of 0.00148, 0.001 and 0.00153 for the one, three and three-day lag analyses. The odds ratios were also quite low reaching only 1.014, 1.0097 and 1.015. The unidirectional post method on the other hand, severely overestimated the parameter with values of 0.00768 and 0.00651 for the unweighted and weighted three-day averages. The one-day analysis was slightly underestimated (0.0045). The odds ratios are similar in that the three-day analyses are overestimating the parameter (1.074 and 1.063, unweighted and weighted) and the one-day analysis is a little low (1.043).

The Breslow method of estimation under the bidirectional design has also underestimated the parameter (0.00295, 0.00422 and 0.00394) and odds ratio (1.028, 1.04 and 1.038) in all three analyses. Efron yielded values in Table 5 that are closer to the 'true' parameter and odds ratio but they too are low. The Efron parameter estimates are 0.00346, 0.00492 and 0.00463, and the estimated odds ratios are 1.033, 1.047 and 1.044.

Exact produced results that are the closest to the parameter of all the bidirectional results. One-day, three-day and three-day lagged analyses produced estimates of 0.00356, 0.00509 and 0.00476 and odds ratios of 1.034, 1.049 and 1.046.

By comparing the values for the estimates and odds ratios to the parameter values, we see that the pre-adjustment in time series analysis gives the best fit with the three-day average ($\hat{\beta}_1$=0.00508, OR=1.0487). With an estimate of 0.00447 and an odds ratio of 1.0494, the one-day average follows close behind the three-day results. With co-

adjustment, the best result falls under the one-day, $PM_{2.5}$ generation (0.00395), but this is quite a difference from 0.00508. This implies that using a pre-adjustment design produced more accurate results than a co-adjustment design (rather, adjusting for temporal trends before analysis is better than adjusting during analysis). This also implies that using a three-day, unweighted average of pollution levels is better than simply using today's pollution level.

All of the unidirectional pre estimates and odds ratios were extremely underestimated, and most of the unidirectional post estimates and odds ratios were extremely overestimated. The one-day analysis with the actual $PM_{2.5}$ level for the generation of admissions provided the closest estimate to the real of the unidirectional designs with a value of 0.00591.

Of the bidirectional designs, the exact method of estimation with the actual $PM_{2.5}$ level for generation of data, under a one-day analysis is the most accurate ($\hat{\beta}_1$=0.00517, OR=1.0497).

The most accurate method for analysis is dependent upon the way that the data are generated. It is expected that the one-day analysis will give more precise parameter estimates than any other analysis under the one-day data generation. After comparing the results from the one- and three-day analyses, it is concluded that the one-day analysis is indeed more precise.

Similarly, the weighted three-day average is expected to give the most accurate results under the lag generation. This is surprisingly not the case here. The simulation yielded results that indicated that the unweighted three-day average is the most accurate. Although the unweighted average is preferable, the weighted average analysis produced

results very similar to those of the unweighted analysis. There is no obvious explanation for this surprising discovery and perhaps should be investigated further at a later date.

3.3 Discussion

Many techniques were utilized for the simulations performed within this chapter. The reasoning behind their usage stems from studies that have been done in the past.

Schwartz *et al.* (1996) made use of the generalized additive model when studying air pollution and its effects on mortality and hospital admissions of the elderly in Cleveland, Ohio. Cakmak *et al.* (1999) used a nonparametric LOESS smoothing function of the day of the study in order to remove temporal cycles from their data for estimating populations threshold concentrations for air pollution related mortality with exposure measurement error.

Lag effects were studied by Burnett *et al.* (1995) in his paper regarding ambient particulate sulfate and its association with hospital admissions for cardiac and respiratory disease. They studied various lag periods from 0 to 3 days and concluded that sulfate levels recorded 1 and 2 days prior to the admission date are positively related to admission rates. Using this as a basis for our decision, a lag of three days was chosen and implemented in our simulation of the Toronto area asthma hospitalization counts.

Burnett *et al.* (1995) also used the 19-day Shumway filter to remove temporal trends from the data. Other papers (Burnett and Krewski, 1994, Schwartz *et al.*, 1996), used a temporal trend filter with positive effects on the results, so it was applied to our simulation as well. Burnett (1994 and 1995) also used a day of the week filter in his papers regarding hospital admissions and mortality.

The bidirectional and time series models seemed to produce results that are closest to the value being estimated, whereas the unidirectional models tend to be biased with an over- or underestimation of the real value. We confirmed the results of Navidi (1998) who concluded that the unidirectional designs are biased because they look solely before or after the date of incidence whereas the bidirectional designs look both before and after, taking more information into consideration. This thesis shows that the estimates from pre-unidirectional are very low compared to the true value whereas the estimates from post-unidirectional are very high.

Bidirectional designs control for time trends (Bateson and Schwartz, 1999) and are arguably a better choice of method than time series. The time series approach is somewhat model dependent since changing the smoothing parameter could drastically change the results. When using a smoothing technique, the smoothing window is subjectively chosen by the researcher. Although some suggest using the autocorrelation plots to minimize the residual autocorrelation, or the Akaike's information criteria (AIC) (Tobías *et al.*, 1999), the length of the span is still completely up to the researcher in the end. The problem with this lack of a standard approach is that different studies may not be directly comparable.

Through all of the averages and different analyses, the one-day, bidirectional, exact method under Model I resulted in estimates that were consistently closest to the true value, 0.00525. From this we can conclude that it is the best method of those studied here for analyzing the hospital admission data and $PM_{2.5}$ levels in Toronto. It is important to note that we do not know exactly how the real data are dependent on the

pollution (whether or not there is a lag), so Model I may not be the most accurate for all types of data.

Table 4. Mean and standard error of $\hat{\beta}_1$, and mean of odds ratio based on 500 simulations generated under $E(y_t) = D_t S_t e^{(\beta_0+\beta_1 x_t)}$, where $\beta_0 = -0.12$, $\beta_1 = 0.00525$, x_t represents the value of the pollutant $PM_{2.5}$ on the t^{th} day, D_t is the day of the week admissions filter and S_t is the Shumway filter for seasonal effects. The relative risk is assumed to be 1.05 and the inter-quartile range is 9.36 $\mu g / m^3$.

Time Series Analysis	Criteria	1 Day Average	3 Day Average
Co-adjustment *Span 93 days*	Mean ($\hat{\beta}_1$)	0.00395	0.00304
	Std. Error	0.00165	0.00212
	Mean (odds ratio)	1.03760	1.02892
Pre-adjustment *Span 93 days*	Mean ($\hat{\beta}_1$)	0.00447	0.00508
	Std. Error	0.00173	0.00235
	Mean (odds ratio)	1.04935	1.04868

Case-Crossover Analysis	Criteria	1 Day Average	3 Day Average
Unidirectional (pre)	Mean ($\hat{\beta}_1$)	0.00271	0.00101
	Std. Error	0.00239	0.00307
	Mean (odds ratio)	1.02577	1.00988
Unidirectional (post)	Mean ($\hat{\beta}_1$)	0.00591	0.00721
	Std. Error	0.00244	0.00318
	Mean (odds ratio)	1.05677	1.05863
Bidirectional (Breslow)	Mean ($\hat{\beta}_1$)	0.00428	0.00408
	Std. Error	0.00207	0.00268
	Mean (odds ratio)	1.04079	1.03885
Bidirectional (Efron)	Mean ($\hat{\beta}_1$)	0.00503	0.00471
	Std. Error	0.00210	0.00271
	Mean (odds ratio)	1.04815	1.04518
Bidirectional (exact)	Mean ($\hat{\beta}_1$)	0.00517	0.00491
	Std. Error	0.00228	0.00295
	Mean (odds ratio)	1.04967	1.04714

Table 5. Mean and standard error of $\hat{\beta}_1$, and mean of odds ratio based on 500 simulations generated under $E(y_t) = D_t S_t e^{(\beta_0 + \beta_1 z_t)}$, where $\beta_0 = -0.12$, $\beta_1 = 0.00525$, z_t represents the value of the three-day lagged pollutant $PM_{2.5}$ on the t^{th} day, D_t is the day of the week admissions filter and S_t is the Shumway filter for seasonal effects. The relative risk is assumed to be 1.05 and the inter-quartile range is 9.36 $\mu g / m^3$.

Time Series Analysis	Criteria	1 Day Average	3 Day Average	3 Day Lag Average
Co-adjustment *Span 93 days*	Mean ($\hat{\beta}_1$)	0.00254	0.00318	0.00334
	Std. Error	0.00168	0.00212	0.00202
	Mean (odds ratio)	1.02407	1.03028	1.03181
Pre-adjustment *Span 93 days*	Mean ($\hat{\beta}_1$)	0.00289	0.00405	0.00410
	Std. Error	0.00175	0.00232	0.00219
	Mean (odds ratio)	1.02734	1.03863	1.03911

Case-Crossover Analysis	Criteria	1 Day Average	3 Day Average	3 Day Lag Average
Unidirectional (pre)	Mean ($\hat{\beta}_1$)	0.00148	0.00100	0.00153
	Std. Error	0.00239	0.00308	0.00292
	Mean (odds ratio)	1.01409	1.00977	1.01475
Unidirectional (post)	Mean ($\hat{\beta}_1$)	0.00450	0.00768	0.00651
	Std. Error	0.00244	0.00317	0.00301
	Mean (odds ratio)	1.04299	1.07447	1.06282
Bidirectional (Breslow)	Mean ($\hat{\beta}_1$)	0.00295	0.00422	0.00394
	Std. Error	0.00208	0.00268	0.00255
	Mean (odds ratio)	1.02803	1.04033	1.03760
Bidirectional (Efron)	Mean ($\hat{\beta}_1$)	0.00346	0.00492	0.00463
	Std. Error	0.00211	0.00272	0.00258
	Mean (odds ratio)	1.03300	1.04730	1.04438
Bidirectional (exact)	Mean ($\hat{\beta}_1$)	0.00356	0.00509	0.00476
	Std. Error	0.00229	0.00295	0.00280
	Mean (odds ratio)	1.03397	1.04897	1.04566

Chapter 4

Conclusions and Limitations

In Chapter 2, the odds ratios obtained under each method of analysis (time series, case-crossover) were not all the same. This creates a problem since it is hard to say which method is giving the most accurate results. A simulation was performed in Chapter 3 to study the different results and decide upon the most appropriate method. It was concluded that the bidirectional design under the exact method of estimation produced results closest to the 'true' values. Knowing this, we now re-examine the results from Chapter 2 and make a more solid conclusion.

It was concluded in Chapter 2 that nitrogen dioxide (NO_2) and carbon monoxide (CO) had the most association with asthma hospitalizations of all the pollutants analyzed. This remains the case when we look solely at the bidirectional results under the exact method of estimation. The odds ratio for NO_2 for the Toronto data is 1.015, and the odds ratio for CO for the Toronto data is 1.001. It was shown that once the covariates of temperature and humidity were added to the model, almost all of the odds ratios decreased slightly, or had virtually no change.

There are some limitations in this study. The daily pollutant values are an average of all four of the pollution monitoring sites for the Toronto area. We are using the average pollution value on a given day to represent the personal exposure for all

people in Toronto. Since these values are not necessarily equivalent to an individual's personal exposure, some error is introduced into our data. Error is also introduced since we are not considering any pollution other than outdoor air pollution. Some pollutants, such as NO_2, are also found indoors, and sometimes at levels that exceed those measured outdoors, thus affecting one's personal exposure (Castellsagué *et al.*, 1995).

Cats, dogs, pollen, and mites can all cause allergic reactions within children which may affect their breathing, or trigger an asthma attack. Cigarette smoke is another factor to be considered. Although most children aged six to twelve do not smoke, the second hand smoke from a parent or guardian changes a child's amount of personal exposure. Where a child lives also affects their level of exposure depending upon whether he lives in the country or the city (Scarlett *et al.*, 1996).

It should be noted that our simulation, as well as our analysis of the Toronto data, used a span of 93 days in the time series analyses and a case-control period of two weeks. By using different spans (186 days, 365 days, *etc.*) different conclusions may arise. The windows of 93 days and two weeks were decided upon based on theories discussed in previous chapters.

Another limitation to this study is with the available data, some of the pollutants have missing values. The data for PM_{10}, $PM_{2.5}$, and NO_2 were essentially complete, whereas CO, O_3 and SO_2 each had some missing values. The missing data were all coded as 0. Since some of the values were actually 0, it is impossible to distinguish a true value of 0 from an assigned value of 0. Thus, all of the pollutants were analyzed in Chapter 2 under the assumption that all 0 values were observed values. This assumption is not all together an unreasonable one. Although SO_2 had 890 data points (18.7%) equal to 0,

some of its non-zero values were quite low, indicating that some of the 890 zeros were observed values. There are very few zero values for CO and O_3 (0.11% and 0.04% respectively). Even if all of the zeros for CO and O_3 are missing values, they should not have a great effect on our results at such low percentages.

Chapter 5

Statistical Techniques

5.1 Autocorrelation Coefficients

An autocorrelation coefficient measures the correlation between observations at different distances apart. In the context of this thesis, they are very useful in that they give some understanding to the model generating the data and, more importantly, aid in the choice of the span, or smoothing window.

The ordinary correlation coefficient is given by,

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \tag{5.1.1}$$

for N pairs of observations on two variables x and y. When this theory is applied to time series, we can measure how strongly successive observations are correlated.

Given N observations from a discrete time series $(x_1,...,x_N)$, $(N-1)$ pairs of observations can be formed $((x_1,x_2),(x_2,x_3),...,(x_{N-1},x_N))$. By treating the first observation of each pair as one variable and the second observation as the other, we can find the correlation coefficient between x_t and x_{t+1} through

$$(5.1.2) \qquad r_1 = \frac{\frac{1}{N-1}\sum_{t=1}^{N-1}(x_t - \bar{x}_{(1)})(x_{t+1} - \bar{x}_{(2)})}{\frac{1}{N-1}\sqrt{\sum_{t=1}^{N-1}(x_t - \bar{x}_{(1)})^2 \sum_{t=1}^{N-1}(x_{t+1} - \bar{x}_{(2)})^2}}$$

where $\bar{x}_{(1)} = \sum_{t=1}^{N-1} x_t/(N-1)$ and $\bar{x}_{(2)} = \sum_{t=2}^{N} x_t/(N-1)$. This is called the autocorrelation coefficient of lag 1 since it measures the correlations between successive observations.

In order to simplify equation (5.1.2) we see that $\bar{x}_{(1)} \approx \bar{x}_{(2)} \approx \bar{x} = \sum_{t=1}^{N} x_t/N$. Instead of taking separate variances in the denominator, we may use the variance of the whole series. Hence, one can define

$$(5.1.3) \qquad r_1 = \frac{\frac{1}{N-1}\sum_{t=1}^{N-1}(x_t - \bar{x})(x_{t+1} - \bar{x})}{\frac{1}{N}\sum_{t=1}^{N}(x_t - \bar{x})^2}.$$

For large N, $(N-1)/N \approx 1$ so we can now simplify (5.1.3) to

$$(5.1.4) \qquad r_1 = \frac{\sum_{t=1}^{N-1}(x_t - \bar{x})(x_{t+1} - \bar{x})}{\sum_{t=1}^{N}(x_t - \bar{x})^2}.$$

Following a similar format, we can find the autocorrelation coefficient at lag k to be,

$$(5.1.5) \qquad r_k = \frac{\frac{1}{N-k}\sum_{t=1}^{N-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\frac{1}{N}\sum_{t=1}^{N}(x_t - \bar{x})^2}.$$

A correlogram is a graph where r_k is plotted against the lag k. It is used to aid in the interpretation of a set of autocorrelation coefficients. Examples of correlograms can be seen in Figures 1 and 2 of this thesis.

Short-term correlation of a stationary series is distinguishable by a correlogram that shows one large correlation coefficient initially, followed by a few more 'large' coefficients that are decreasing in value. It is easy to see from Figures 1 and 2 that the Toronto data fits this description and thus can be considered as short-term correlation of a stationary series.

As was mentioned previously, the appropriate spans for each pollutant were determined by analyzing the autocorrelation functions (ACF) of the data. The best span for a particular pollutant is the one whose ACF plot (correlogram) shows that as the number of lags increase, the smaller (closer to 0) the coefficients become.

5.2 The Generalized Additive Model

The principal model used for the time series analysis was the generalized additive model. The generalized additive models (GAM) are extensions of the generalized linear models (GLM). The GLM uses a linear function of the parameters in the model whereas the GAM fits nonparametric functions for estimating the response and predictor relationships. Smoothing processes are used to estimate these functions from the given data and each variable must be analyzed according to the smoothing parameter most appropriate for it. For our purpose, the predictors are the air pollutants and seasonal factors, and the response is the daily count of hospitalizations for asthma. The spans that were mentioned previously are being used as the smoothing parameters.

The basic form of GAM is

$$(5.2.1) \qquad g(E(Y \mid x)) = \alpha + \sum_{i=1}^{n} f_i(x_i)$$

where g is the link function, α is the constant intercept, f_i is the nonparametric function linking the transformed mean response to the i^{th} predictor, and x_i is the i^{th} predictor variable. A link function relates the canonical parameters of the specified distribution family to the linear predictor. In this thesis the link used is Poisson, thus a Poisson regression model is fit. This link was chosen because we assume the daily admissions to follow a Poisson distribution.

The smoothing operation LOESS was used along with GAM, to smooth the temporal trends both in the analysis of the Toronto data and in the simulation. In S-Plus, the *gam* function is able to call *lo* to fit a locally weighted least squares regression. Cakmak *et al.* (1999) also made use of the LOESS regression smoother in a study relating air pollution and mortality.

5.3 Case-Crossover Analysis

As stated in Chapter 2, case-crossover is somewhat of a modified version of the case-control technique. Case-control matches a set of cases to a separate set of controls. For the Toronto data, a case is defined as a date with an asthma hospitalization whereas a control is defined as a date without an asthma hospitalization. With case-crossover, the cases use themselves as the controls thus reducing the bias which occurs from the case-control design.

Within this thesis, two different case-crossover designs were used. The first one, unidirectional, looks at the case and compares it with a control dated either before (pre) or after (post) it. The second one, bidirectional looks at the case and compares it with two controls, one dated before and one dated after the case date.

Logistic regression was used in the analysis of the case-crossover designs. Likelihood estimates are being used here to develop a model from the general case-control model to which logistic regression can be applied.

Using the general case-control model, the conditional likelihood for the first n_1 $\underset{\sim}{x}$'s are cases can be expressed as follows:

$$(5.3.1) \qquad l_k(\beta) = \frac{\prod_{i=1}^{n_{1k}} P(\underset{\sim}{x}_i \mid y=1) \prod_{i=n_{1k}+1}^{n_k} P(\underset{\sim}{x}_i \mid y=0)}{\sum_j \left\{ \prod_{i_j=1}^{n_{1k}} P(\underset{\sim}{x}_{ji_j} \mid y=1) \prod_{i_j=n_{1k}+1}^{n_k} P(\underset{\sim}{x}_{ji_j} \mid y=0) \right\}},$$

where $y=1$ refers to the cases and $y=0$ refers to the controls in each stratum, and $\underset{\sim}{x}$ is the set of covariates or exposure variables. In the model the stratum is k for $k=1,2,...K$ (K=4629, total number of stratum) with n_{0k} controls, and n_{1k} cases. Of the $n_k = n_{1k} + n_{0k}$ total individuals, there are $\binom{n_k}{n_{1k}}$ possibilities for n_{1k} of them to receive case status. The sum in the denominator is over all the $\binom{n_k}{n_{1k}}$ ways of dividing the numbers into groups. The probability of the observed outcome as it relates to the probability of possibilities for n_{1k} is the conditional likelihood (5.3.1) for the k^{th} stratum. The subscript j will denote a possible assignment of n_{1k}. Keeping all the cases together and all of the controls together, we let $i=1$ to n_k, where 1 to n_k refer to the cases, and $n_{1k}+1$ to n_k refer to the

controls. The index i corresponds to observed data whereas i_j is for the j^{th} assignment possibility.

The logistic regression model is defined as

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}. \qquad (5.3.2)$$

Using a logit, we can transform the logistic regression model into a linear regression model. The logit is defined as

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x. \qquad (5.3.3)$$

Logit models use the explanatory variables to express the log odds of being in a specific category of the response variable. In terms of this thesis, the logit is describing the log odds of being admitted to the hospital for asthma by the pollution in the air near the time of admission.

If we assume that β_k indicates the contribution to the logit made by the constant terms within stratification variables, then $g_k(\underset{\sim}{x}) = \beta_k + \underset{\sim}{\beta}'\underset{\sim}{x}$ represents the logit in the k^{th} stratum, where $\underset{\sim}{x}$ is a vector of pollutant values, and $\frac{e^{\beta_k+\beta_{\underset{\sim}{x}}}}{1+e^{\beta_k+\beta_{\underset{\sim}{x}}}}$ is the logistic function. By substituting this logistic function into (5.3.1) we see that,

$$l_k(\beta) = \frac{\prod_{i=1}^{n_{1k}} e^{\beta' \underset{\sim}{x}_i}}{\sum_j \prod_{i_j=1}^{n_{1k}} e^{\beta' \underset{\sim}{x}_{ji_j}}} \qquad (5.3.4)$$

after simplification (both the numerator and the denominator contain the term $e^{\beta_k}/(1+e^{\beta_k+\beta_{\underset{\sim}{x}}})$ thus are cancelled).

By multiplying (5.3.4) over K strata we can obtain the conditional likelihood, $l(\beta) = \prod_{k=1}^{K} l_k(\beta)$. Maximizing $l(\beta)$, yields the conditional maximum likelihood estimate of β.

On the other hand, in case-crossover with 1-1 matching, the general conditional logistic model is,

$$(5.3.5) \qquad l_k(\beta) = \frac{e^{\beta' x_{1k}}}{e^{\beta' x_{1k}} + e^{\beta' x_{0k}}}$$

for the k^{th} strata where $\underset{\sim}{x}_{1k}$ is the vector corresponding to the k^{th} case, and $\underset{\sim}{x}_{0k}$ is the vector corresponding to the k^{th} control. Dividing both the numerator and the denominator by $e^{\beta x_{0k}}$ transforms (5.3.3) into

$$(5.3.6) \qquad l_k(\beta) = \frac{e^{\beta'(x_{1k} - x_{0k})}}{1 + e^{\beta'(x_{1k} - x_{0k})}}$$

If $x_k^* = x_{1k} - x_{0k}$, then (5.3.6) is simply a logistic regression model with β_0, the constant term, equal to 0.

Now, by using the number of case-control pairs as the sample size, the differences x_k^* as the covariates, and omitting the constant term, the conditional maximum likelihood estimate can be found via standard logistic regression. Also, all response variables (y_k) values must be coded as 1.

5.4 Cox Proportional Hazards Model

The Toronto data was stratified into two groups and labeled 1 if it refers to a case and 0 if it refers to a control. This stratified data set was analyzed using the Cox

Proportional Hazards model (CoxPH). CoxPH is the most widely used model for survival analysis since it works well with both discrete and continuous data. Also, specific probability distributions need not be chosen to represent survival times as is necessary with parametric regression models.

Before exploring the inner workings of CoxPH, the term hazard function should be defined. A hazard function, in the continuous case is the individual's rate of failure at $T=t$ conditional upon the individual's survival to time t. Our data uses admission dates as incidences or 'failures' and thus is considered discrete. Since T is a discrete random variable in this case, taking values $x_1 < x_2 < \ldots$, its associated probability function is

$$f(x_i) = P(T = x_i), \quad i = 1,2,... \tag{5.4.1}$$

The hazard function for discrete data is defined as the conditional probability of failure at x_j given that he has survived up to x_j, $\lambda_j = P(t = x_j \mid T \geq x_j)$.

The hazard function for a particular individual in the j^{th} stratum is represented by

$$\lambda_j(t;z) = \lambda_{0j}(t)e^{z\beta} \tag{5.4.2}$$

where λ_0 is the arbitrary, unspecified hazard function and $j = 1,2,...r$ represents the stratum. z is a vector containing covariates at time t for the i^{th} individual, and β is the regression parameter vector, less the intercept (it is absorbed into λ_0). This is Cox's proportionate hazard.

The regression parameter vector β must now be estimated using the hazard function (5.4.2). This estimation can be done using the full likelihood (estimating β and nuisance parameters jointly) or by eliminating the nuisance parameters by considering an appropriate conditional distribution (partial likelihood). The full and partial likelihoods

contain most of the same information about β for regression with censored data, so for the sake of simplicity, the partial likelihood is the method used here.

Consider that if at time $t_{(i)}-0$, $R(t_{(i)})$ is the set of individuals at risk, then

$$\frac{\lambda(t_{(i)};z_{(i)})}{\sum_{l\in R(t_{(i)})}\lambda(t_{(i)};z_l)}=\frac{e^{(z_{(i)}\beta)}}{\sum_{l\in R(t_{(i)})}e^{(z_l\beta)}} \quad \text{for } i=1,...k \tag{5.4.3}$$

is the conditional probability that individual i fails (is admitted to the hospital) at time $t_{(i)}$. This model also assumes that only one admission occurs at $t_{(i)}$ and the individuals of $R(t_{(i)})$ are in fact at risk. If there is no admission between time $t_{(i-1)}$ and $t_{(i)}$, then no further information can be gathered on β since $\lambda_0(t)$ is unspecified.

Taking the product of the conditional probabilities of admissions results in,

$$L(\beta)=\prod_{i=1}^{k}\frac{e^{(z_{(i)}\beta)}}{\sum_{l\in R(t_{(i)})}e^{(z_l\beta)}}. \tag{5.4.4}$$

This is simply the product of the conditional likelihood in (5.3.4), which is why the CoxPH routine is used to solve (5.3.4). Cox (1972) has shown that these estimates are asymptotically normally distributed as sample size increases.

The problem with this model is that it assumes that there is only one admission per day. Looking at the data we see that this is not the case. Some days have as many as 11 admissions. Having more than one incident occurring at the same time (more than one admission per day) is called a tie. By using a similar argument with the discrete logistic model, a partial likelihood can be attained which accounts for ties in the data (more than one admission per day). Where $\lambda_d(t)$ is an unspecified discrete hazard,

$$\frac{\lambda_d(t)\,dt * e^{(z\beta)}}{1-\lambda_d(t)\,dt} \tag{5.4.5}$$

is the hazard relationship for the partial likelihood which can be obtained by applying the discrete logistic model. From this, the conditional probability that the d_i failures should be those observed given the risk set and multiplicity of d_i, is the i^{th} term of the product of the hazard relationship,

$$\prod_{i=1}^{k} \left(\frac{e^{(s_i \beta)}}{\sum_{l \in R_{d_i}(t_{(i)})} e^{\sum_{j=1}^{d_i} z_{lj} \beta}} \right) \qquad (5.4.6)$$

with s_i being the sum of the covariates in z of the d_i admissions on day $t_{(i)}$. The set of all subsets of d_i individuals selected from $R(t_{(i)})$, without replacement, is represented by $R_{d_i}(t_{(i)})$.

The partial likelihood, (5.4.6) is not consistent in its estimation of β when ties arise by the grouping of continuous failure times. This is explained by the fact that (5.4.6) emerges from the discrete model and so estimates β in that model. Since the discrete model does not arise due to the grouping of continuous failure times, the two parameters will not be identical in their interpretations.

The computer software S-Plus allows us three choices for the method of approximation: Breslow, Efron, and Exact. Breslow suggests using the marginal likelihood approach in order to estimate β. This idea is fine, but the result becomes computationally awkward when there are a large number of ties at any failure time. Efron proposed using the Kaplan-Meier estimator when dealing with ties. This is closely related to Breslow (1974) but Efron's estimation gives a closer approximation to the real. The exact method is the best choice of the three in terms of producing good estimators for β, but when the data contain a lot of ties it is extremely time consuming. The exact

function in S-Plus uses the partial likelihood method, as was elaborated upon earlier, to approximate. Each of these methods was utilized in the case-crossover simulation.

References

Allison. (1995). Survival Analysis Using the SAS System A Practical Guide. *SAS Institute Inc.*

Bai, Kenney and Von Gorder. (2001). Respiratory Disease in Canada. *Population and publication Health Branch (PPHB), Available* http://www.hc-sc.gc.ca/pphb-dgspsp/publicat/rdc-mrc01/

Bascom, Bromberg, Costa, Bevlin, Dockery, Frampton, *et al.* (1996). Health effects of outdoor air pollution. *American Journal of Respiratory Critical Care Medicine*, 153, 3-55

Bates, Baker-Anderson and Sizto. (1990). Asthma attack periodicity: A study of hospital emergency visits in Vancouver. *Environmental Research*, 51, 51-70.

Bates, Sizto. (1987). Air pollution and hospital admissions in Southern Ontario: The acid summer haze effect. *Environmental Research*, 43, 317-331.

Bateson and Schwartz. (1999). Control for seasonal variation and time trend in case-crossover studies of acute effects of environmental exposures. *Epidemiology*, 10, 539-544.

Breslow. (1972). Contribution to the discussion of D. R. Cox (1972). *Journal of the Royal Statistical Society*, Series B, 34, 216-217.

Breslow. (1974). Covariance analysis of censored data. *Annals of Statistics*, 2, 437-453

Breslow and Day. (1980). Statistical Methods in Cancer Research. *International Agency for Research on Cancer Scientific Publications No. 32*

Burnett, Dales, Krewski, Vincent, Dann, and Brook. (1995). Associations between ambient particulate sulfate and admissions to Ontario hospitals for cardiac and respiratory diseases. *American Journal of Epidemiology*, 142, 15-22.

Burnett, Dales, Razenne, Krewski, Summers, Roberts, Raad-Young, Dann, and Brook.

(1994). Effects of low ambient levels of ozone and sulfates on the frequency of respiratory admission to Ontario hospitals. *Environmental Research*, 65, 172-194.

Burnett and Krewski. (1994). Air pollutions effects on hospital admission rates: A random effects modeling approach. *The Canadian Journal of Statistics*, 22, 441-458.

Burnett, Smith-Doiron, Stieb, and Cakmak. (1999). Effects of particulate and gaseous air pollution on cardiorespiratory hospitalizations. *Archives of Environmental Health*, 54, 130-139.

Cakmak, Burnett, Krewski. (1999). Methods for detecting and estimating population threshold concentrations for air pollution-related mortality with exposure measurement error. *Risk Analysis*, 19, 487-496.

Castellsagué, Sunyer, Sáez, and Antó. (1995). Short-term addociation between air pollutions and emergency room visits for asthma in Barcelona. *Thorax*, 50, 1051-1056.

Chatfield. (1984). The Analysis of Time Series An Introduction. *Chapman and Hall Ltd.*

Cox. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society,* Series B, 34, 187-220.

D'Amato (G), Liccardi, Russo, D'Amato(M). (2000). On the interrelationship between outdoor air pollution and respiratory allergy. *Aerobiologia*, 16, 1-6.

Devalia, Rusznak, Wang, Khair, Muntasir, Abdelaziz, Calderón, Davies. (1996). Air pollutants and respiratory hypersensitivity. *Toxicology Letters*, 86, 169-176.

Efron. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72, 557-565

Elandt-Johnson, Johnson. (1980). Survival Models and Data Analysis. *John Wiley and Sons Inc.*

Hosmer and Lemeshow. (1989). Applied Logistic Regression. *John Wiley and Sons Inc.*

Kalbfleisch and Prentice. (1980). The Statistical Analysis of Failure Time Data. *John Wiley and Sons Inc.*

Kendall and Stewart. (1966). The Advanced Theory of Statistics.

Koenig. (1999). Mortality and air pollution: Associations persist with continued advances in research methodology. *Environmental Health Perspectives*, 107, 613-615.

Lawless. (1982). Statistical Models & Methods For Lifetime Data. *John Wiley and Sons Inc.*

Lee and Schwartz. (1999). Reanalysis of the effects of air pollutions on daily mortality in Seoul, Korea: A case-crossover design. *Environmental Health Perspectives*, 107, 633-636.

Linn, Shamoo, Anderson, Whynot, Avol, and Hackney. (1985). Effects of heat and humidity on the responses of exercising asthmatics to sulfur dioxide exposure. *American Review of Respiratory Disease*, 131, 221-225.

Mackenbach, Looman, Kunst. (1993). Air pollution, lagged effects of temperature, and mortality; The Netherlands 1979-87. *Journal of Epidemiology and Community Health*, 47, 121-126

Maclure. (1991). The case-crossover design: A method for studying transient effects on the risk of acute events. *American Journal of Epidemiology*, 133, 144-153.

Navidi. (1998). Bidirectional case-crossover designs for exposures with time trends. *Biometrics*, 54, 596-605.

Neas, Schwartz, Dockery. (1999). A case-crossover analysis of air pollutions and mortality in Philadelphia. *Environmental Health Perspectives*, 107, 629-632.

Nicolai. (1999). Air pollution and respiratory disease in children: What is the clinically relevant impact? *Pediatric Pulmonology*, Supplement 18, 9-13.

Scarlett, Abbott, Peacock, Strachan, Anderson. (1996). Acute effects of summer air pollution on respiratory function in primary school children in southern England. *Thorax*, 51, 1109-1114.

Schwartz and Dockery. (1992). Particulate air pollution and daily mortality in Steubenville, Ohio. *American Journal of Epidemiology*, 135, 12-19.

Schwartz and Morris. (1995). Air pollutions and hospical admissions for cardiovascular disease in Detroit, Michigan. *American Journal of Epidemiology*, 142, 23-35.

Schwartz, Slater, Larson, Pierson, Koenig. (1993). Particulate air pollutions and hospital emergency room visits for asthma in Seattle. *American Review of Respiratory Disease*, 147, 826-831.

Schwartz, Spix, Touloumi, Bachárová, Barumamdzadeh, le Tertre, Piekarksi, Ponce de León, Pönkä, Rossi, Sáez, and Schouten. (1996). Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. *Journal of Epidemiology and Community Health*, 50(Suppl), S3-S11.

Shumway, Tai (R.), Tai (L.), and Pawitan. (1983). Statistical analysis of daily London mortality and associated weather and pollution effects. *Technical Report 53, Division of Statistics, University of California, Davis.*

Therneau, Bates, Pinheiro, Ripley, Venables. (1999). S-Plus 2000 Modern Statistics and Advanced Graphics. *MathSoft Inc.*

Sunyer, Castellsagué, Sáez, Tobías, and Antó. (1996). Air pollutions and mortality in Barcelona. *Journal of Epidemiology and Community Health,* 50 (Suppl.), S76-S80.

Sunyer, Spix, Quénel, Ponce de León, Pönkä, Barumandzadeh, Tounloumi, Bachárová, Woytyniak, Vonk, Bisanti, Schwartz, Katsouyanni. (1997). Urban air pollution and emergency admissions for asthma in four European cities: the APHEA Project. *Thorax,* 52, 760-765.

Tobías, Campbell, Sáez. (1999). Modelling asthma epidemics on the relationship between air pollution and asthma emergency visits in Barcelona, Spain. *European Journal of Epidemiology,* 15, 799-708.

U.S. EPA Office of Air Quality Planning and Standards. (May 31, 2002). National Ambient Air Quality Standards (ANNQS). *Available* http://www.epa.gov/airs/criteria.html

Vita Auctoris

Name:	**Abby Livingston**
Date of Birth:	**1979**
Place of Birth:	**Leamington, Ontario, Canada**
Educations:	**B.A. in Mathematics and Statistics, 2001** **University of Windsor, Canada** **M.S. in Statistics, 2002** **University of Windsor, Canada**